



UNIVERSITAT^{DE}
BARCELONA

Re-definition of non-small cell lung cancer transcriptional subtypes using integrative bioinformatics approaches

Sara Hijazo Pechero



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution 4.0. Spain License.**

UNIVERSITAT DE BARCELONA

FACULTAT DE MEDICINA I CIÈNCIES DE LA SALUT


Tesi realitzada a l'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

PROGRAMA DE DOCTORAT DE BIOMEDICINA

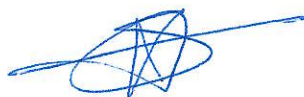
Re-definition of non-small cell lung cancer transcriptional subtypes using integrative bioinformatics approaches

Memòria presentada per **Sara Hijazo Pechero** per optar al títol de Doctor per la Universitat de Barcelona

Doctorand: Sara Hijazo Pechero



Directors: Xavier Solé i Ernest Nadal



Tutor: Víctor Raúl Moreno Aguado



Barcelona, 2024

AGRAÏMENTS

Fa uns mesos vaig començar a fer escalda un parell de dies a la setmana després de la feina. Recordo que al principi em costava bastant millorar i pensava que era perquè em faltava condició física. Tanmateix, amb el temps vaig entendre que la clau d'aquest esport és més aviat entendre quina és la millor combinació de moviments i peces que s'han de fer servir per arribar al cim sense utilitzar tanta força i confiar molt en un mateix i en el procés. A més, també és important aprendre dels companys, i poder aixoplugar-te en ells quan portes tres hores intentant pujar una paret i no surt. En aquests moments, les paraules d'ànim i una abraçada és el que fa que acabis la classe amb ganes de tornar el proper dia i tornar a intentar-ho. I al final surt.

Mirat amb perspectiva, desenvolupar una tesi doctoral no dista tant d'aprendre a escalar. És una carrera de fons, que no hagués estat possible sense el suport de moltes persones, que de diferents formes han contribuït en aquest procés. Aquesta secció és per vosaltres, gràcies a TOTS i TOTES.

En primer lloc, vull agrair als meus directors Xavier Solé i Ernest Nadal, l'oportunitat que em van donar al proposar-me treballar amb ells i el suport constant al llarg de tot el procés. M'agradaria ressaltar que em sento afortunada perquè a part de ser dos persones amb molts coneixements científics en el seu camp i de les quals he après moltíssim, també es pot parlar amb ells a nivell més personal i sempre han estat un

suport molt gran durant aquests moments d'inseguretat que a vegades tenim els estudiants de doctorat. Gràcies per sempre confiar en mi i la meva capacitat per treure la feina endavant. La Cristina Muñoz-Pinedo va aparèixer més tard en el transcurs d'aquesta aventura. Amb ella em vaig adonar que m'agrada molt el metabolisme i l'impacte de l'estrès metabòlic per la progressió tumoral. La Cris és una persona amb qui m'agrada molt pensar i debatre sobre noves hipòtesis i possibles projectes, crec que fem bon equip i a vegades ens retroalimentem positivament. Gràcies per ser sempre tan propera i obrir aquest espai amb nosaltres, crec que parlo per tots els membres del lab si dic que aprenem molt de tu cada dia.

Als meus amics i companys de la planta 2 de l'IDIBELL (Ania, María, Sandra i Raúl), gràcies per les estones que hem passat plegats a la feina i pels moments de disbauxa a la Flama. Ania, gràcies per la teva paciència i per haver estat una molt bona mentora en tot aquest procés. Ets una persona sempre disposada a ajudar a tothom, no només en l'àmbit laboral sinó també en el personal, és una cosa que sempre he admirat de tu. Gràcies per fer-lo sempre tot una mica més fàcil.

A la Carlota li vull agrair que m'hagi acompanyat des de l'inici d'aquest viatge. Has estat un suport molt important, sempre disposada a escoltar-me i animar-me a continuar en els moments que potser em faltaven forces. Saps que has estat i ets una persona molt especial en la meua vida. Espero que, tot i que d'una altra forma, mantinguem el contacte i ens

puguem aportar coses positives. També vull incloure aquí un agraïment especial als teus pares, Carmen i Nico, que sempre s'han preocupat per mi i m'han acollit com si fos filla seva, gràcies de debò per la vostra estima incondicional.

A tots els companys i amics de laboratori: Sílvia, Didac, Carla, Lidia, Fedra, Rut, Felipe, Mabel, Isha, Joaquim, Mireia, Noelia, Anna, David, Emma. Sou un grup molt especial, gràcies per acollir-me tan bé des del principi, entre tots feu que anar cada dia al lab sigui una mica més divertit. Sílvia, gràcies per ser la millor companya de taula i per ser allà sempre que ho he necessitat, has estat un aixopluc dins i fora del lab. És una sort compartir temps amb una persona amb tanta llum per dins. Didac, fa poc temps que ens coneixem però des del principi vam encaixar i connectar súper bé. Ets una persona súper positiva i espero que continuem vivint experiències junts, començant per Monegros i després el que vingui! A la Carla li vull agrair els dimecres de biblioteca, els vermuts al sol (i a l'ombra), les passejades sense rumb i en general els plans random. Has estat tot un descobriment i espero seguir compartint estones amb tu. Com diria el Felipe: “¡No cambien!”.

Als meus amics i amigues de fora de la feina: Andrea, Clara, Gerard, Melissa, Irene, Eloi...gràcies per ser allà sempre que ho he necessitat. Compartir temps amb vosaltres és sempre un regal. Aquesta tesi no hagués estat possible sense el vostre suport incondicional.

Por último, pero no menos importante me gustaría dedicar esta tesis a mi familia: mamá, Chema, Laura, Irene, Isaac, Raquel, Paula y el pequeño Lucas. Gracias por siempre estar presentes de forma incondicional y por confiar siempre en mí. Soy afortunada de tener una familia como la nuestra. Os quiero mucho a todos y, en especial, esta tesis va para vosotros.

Barcelona, Gener de 2024.

Table of Contents

1. SUMMARY	1
2. INTRODUCTION	5
2.1 Current clinical management of non-small cell lung cancer: the need to dive deeper	5
2.1.1 Non-small cell lung cancer epidemiology	7
2.1.2 Towards an increasingly accurate non-small cell lung cancer classification: from histological to molecular subtyping	11
2.1.3 Immunotherapy in non-small cell lung cancer: the revolutionary weapon.....	13
2.1.4 Limitations of the current management framework of non-small cell lung cancer patients.....	16
2.2 Transcriptional profiling as a tool for tackling complexity and molecular heterogeneity in lung adenocarcinoma and lung squamous cell carcinoma.....	18
2.2.1 Lung adenocarcinoma transcriptional subtypes.....	19
2.2.2 Lung squamous cell carcinoma transcriptional subtypes	25
2.2.3 Clinical relevance of lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based classifications.....	30
2.3 Potential applications of transcriptional profiling in the clinical setting: adding another layer of information	33

2.3.1 Patients lacking actionable driver alterations: opening a window for new therapeutic strategies.....	34
2.3.2 Anticipate and overcome potential treatment resistance mechanisms	37
2.3.3 Patient selection for immunotherapy treatment regimens	40
2.3.4 Patients harboring targetable genomic alterations: do all patients follow the same course upon treatment?	45
2.4 Integration of drug sensitivity data from large cancer cell lines pharmacogenomic studies for specific treatment strategies identification	48
2.4.1 Large cancer cell lines pharmacogenomic projects: CCLE, GDSC, CTRP, PRISM.....	49
2.4.2 Large cancer cell lines pharmacogenomic projects as a tool for bringing precision medicine to patients with cancer	51
3. HYPOTHESIS	55
4. OBJECTIVES.....	57
4.1 General objectives	57
4.2 Specific objectives	57
5. RESULTS.....	59
5.1 Publications	59
5.2 Global results summary.....	60

5.2.1 Directors' report.....	60
5.2.2 Article 1: Transcriptional profiling of molecular pathways allows for the definition of lung squamous cell carcinoma molecular subtypes with specific vulnerabilities	61
5.2.3 Article 2: Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications	63
6. DISCUSSION	66
6.1 Transcriptional profiling of molecular pathways yields a robust classification of lung adenocarcinoma and lung squamous cell carcinoma	66
6.2 Pathway transcriptional profiling-based classification correlates and further subdivides widely accepted Wilkerson et al.'s mRNA-based subtypes	69
6.3 Integration of transcriptomic and genomic data could improve current NSCLC patient stratification in patients with driver positive lung adenocarcinoma	72
6.4 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes show differential genome instability features	73
6.5 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes display specific therapeutic vulnerabilities	74

6.6 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes display different immune landscapes with potential therapeutic implications	76
6.7 Limitations of the present work and future perspectives	79
7. CONCLUSIONS	82

Table of Figures

Figure 1. Estimated number of new cancer cases worldwide during 2020.	8
Figure 2. 5-year relative survival by stage at diagnosis.	9
Figure 3. Estimated number of (A) incident cases and (B) deaths worldwide by 2040.	10
Figure 4. Non-small cell lung cancer subtyping evolution: from histological to molecular stratification.	12
Figure 5. Transcriptional-based lung adenocarcinoma molecular subtypes.	21
Figure 6. Transcriptional-based lung squamous cell carcinoma molecular subtypes.	27
Figure 7. NSCLC transcriptional subtypes gene expression signatures overlap.	31
Figure 8. UpSet plot describing the number of treatments identified using different data types, alone (indicated by filled circles) or in combination (indicated by lines	

connecting different data types represented by filled circles).....	36
Figure 9. Mechanisms of resistance to targeted therapies.	38
Figure 10. TCGA tumor immunophenotypes.	42
Figure 11. Evaluation of the predictive usefulness of the 18-gene T cell–inflamed GEP compared to PD-L1 IHC in predicting response to pembrolizumab in a PD-L1–unselected cohort of 96 patients with HNSCC from KEYNOTE-012.....	43
Figure 12. Predictive abilities of M1 signature, peripheral T cell signature, PD-L1 expression, tumor infiltrating lymphocytes (TIL), and tumor mutation burden (TMB)...	44
Figure 13. DepMap project data and aims.....	50
Figure 14. Tumor types by their cancer cell line representation in the dependency screens dataset.	52
Figure 15. Links between pathway transcriptional profiling-based subtypes and Wilkerson et al.’s mRNA-based subtypes.	71

List of Acronyms

ALK	Anaplastic lymphoma kinase
B-I	Basal-inclusive
BRAF	V-RAF murine sarcoma viral oncogene homolog B
CCLE	Cancer Cell Line Encyclopedia
CCLs	Cancer cell lines
CNA	Copy number alterations
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CTLA-4	Cytotoxic T-Lymphocyte Antigen 4
CTRP	Cancer Therapeutics Response Portal
DDR	DNA damage repair
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
EMT	Epithelial mesenchymal transition
EMT-E	Epithelial mesenchymal transition-enriched
ERBB2	V-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
FF	Fresh frozen
FFPE	Formalin fixed paraffin embedded
GDSC	Genomics of Drug Sensitivity in Cancer
GSVA	Gene set enrichment analysis
IARC	International Agency for Research on Cancer
ICB	Immune checkpoint blocker
ICI	Immune checkpoint inhibitor
I-S	Inflamed-secretory
KEAP1	Kelch-like ECH-associated protein 1
KRAS	Kirsten rat sarcoma viral oncogene
LC	Lung cancer
LCC	Large cell carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MET	Mesenchymal-epithelial transition factor
NEF2L2	Nuclear factor erythroid 2-related factor 2

NF1	Neurofibromin 1
NGS	Next generation sequencing
NMSC	Non-melanoma skin cancer
NOS	Not otherwise specified
NSCLC	Non-small cell lung cancer
NTRK	Neurotrophic tyrosine receptor kinase
PD-1	Programmed death-receptor 1
PD-L1	Programmed death-ligand 1
PI	Proximal inflammatory
PP	Proximal proliferative
P-P	Proliferative-primitive
PRISM	Profiling Relative Inhibition Simultaneously in Mixtures
PTEN	Phosphatase and tensin homolog
RB1	Retinoblastoma 1 gene
RET	Ret proto-oncogene
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROS1	C-Ros oncogene 1
SCLC	Small cell lung cancer
SNV	Single nucleotide variant
SOX2	SRY-Box Transcription Factor 2
STK11	Serine/Threonine Kinase 11
TCGA	The Cancer Genome Atlas
TKI	Tyrosine kinase inhibitor
TMB	Tumor mutational burden
TME	Tumor immune microenvironment
TP53	Tumor protein p53
TP63	Tumor protein p63
TRU	Terminal respiratory unit
WHO	World Health Organization
WNT	Wingless-type MMTV integration site family genes
WGTA	Whole genome and transcriptome analyses

1. SUMMARY

Recent technological advances and the utilization of high-throughput DNA and RNA sequencing analyses have increased our understanding of cancer diseases, including non-small cell lung cancer. For instance, genomic characterization has allowed the identification of some gene alterations which can be targeted with specific drug compounds. However, detection of genomic alterations does not fully recapitulate the heterogeneity of the disease and it sometimes fail to predict the subset of patients that most benefit from chemo- or immunotherapy. In this context, transcriptional profiling has emerged as a promising tool for patient selection and treatment guidance. Comprehensive evaluation of the expression level of pathways and genes involved in tumor progression and immune response has demonstrated to predict clinical benefit potentially better than single agents such as PD-L1 or tumor mutational burden in the case of immunotherapy regimens.

This study aimed to establish a robust classification of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) tumors, respectively, based on the transcriptional profiling of 50 landmark molecular pathways. This work also aimed to characterize the subtypes at different levels and to identify potential specific vulnerabilities and drug candidates. Thus, in this thesis we present a new tumor classification framework based on the expression profiling of 50 signaling

pathways, which is more robust than the use of single gene expression levels, prone to multiple sources of variability. Also, this approach allows for the integration of different gene expression datasets, significantly increasing the sample size and the amount of molecular heterogeneity that can be considered. In fact, to our knowledge, no previous LUAD or LUSC classification has been derived from such a large sample size. Moreover, we have provided a way for drug prioritization, based on the molecular characteristics of each subtype and the integration of huge cancer cell lines pharmacogenomics projects. In the end, this work could lay the foundation for improving patient stratification beyond genomics and single biomarkers and pave the way for more personalized treatment avenues in non-small cell lung cancer.

RESUM

Els avenços tecnològics recents i la utilització d'anàlisis de seqüenciació d'ADN i ARN d'alt rendiment han augmentat la nostra comprensió de les malalties del càncer, inclòs el càncer de pulmó de cèl·lula no petita. Per exemple, la caracterització genòmica ha permès identificar algunes alteracions genètiques que poden ser tractades amb fàrmacs específics. No obstant això, la detecció d'alteracions genòmiques no recapitula del tot l'heterogeneïtat de la malaltia i a vegades no aconsegueix predir el subconjunt de pacients que més es beneficiarien del tractament amb quimioteràpia o immunoteràpia. En aquest context, l'anàlisi dels perfils transcripcionals ha sorgit com una eina prometedora per a la selecció del pacient i l'orientació del tractament. Per exemple, l'avaluació del nivell d'expressió de vies i gens implicats en la progressió tumoral i la resposta immunitària ha demostrat predir el benefici clínic de millor manera que marcadors individuals com el PD-L1 o la càrrega mutacional del tumor en el cas de la resposta a immunoteràpia.

Aquest estudi tenia com a objectiu establir una classificació robusta dels tumors d'adenocarcinoma pulmonar i carcinoma de cèl·lules escamoses pulmonars, respectivament. Aquest treball també tenia com a objectiu caracteritzar els subtipus a diferents nivells i identificar possibles vulnerabilitats específiques.

Així, en aquesta tesi presentem un nou marc de classificació tumoral basat en el perfil d'expressió de 50 vies de senyalització, que és més robust que l'ús de nivells d'expressió de gens individuals, propens a múltiples fonts de variabilitat. A més, aquest enfocament permet la integració de diferents conjunts de dades d'expressió gènica, augmentant significativament la mida de la mostra i la quantitat d'heterogeneïtat molecular que es pot considerar. De fet, segons el nostre coneixement, no s'ha obtingut cap classificació prèvia d' adenocarcinoma pulmonar o carcinoma de cèl·lules escamoses pulmonars a partir d'una quantitat de mostra tan gran. D'altra banda, hem donat una aproximació per a la prioritització de fàrmacs, basada en les característiques moleculars de cada subtipus i la integració de grans projectes de farmacogenòmica de línies cel·lulars de càncer. Al final, aquest treball podria establir les bases per millorar l'estratificació del pacient més enllà de la genòmica i els biomarcadors individuals i aplanar el camí per a opcions de tractament més personalitzades en càncer de pulmó de cèl·lula no petita.

2. INTRODUCTION

2.1 Current clinical management of non-small cell lung cancer: the need to dive deeper

Non-small cell lung cancer (NSCLC) is a complex and highly heterogeneous disease. However, this complexity does not only rely on the tumor's histological and morphological characteristics. For instance, tumors with the same histology (e.g., lung adenocarcinoma (LUAD)) and pathological stage do not necessarily follow the same clinical course or exhibit equivalent responses upon the same therapeutic strategy (1). The breakthrough discovery of Epidermal Growth Factor Receptor (*EGFR*) mutations two decades ago led to a paradigm shift in the clinical management and launched the era of personalized medicine in advanced NSCLC. The correlation of clinical response to EGFR tyrosine kinase inhibitors and the identification of those actionable alterations led to the refinement of the existing treatment algorithm, previously based only on histopathological features (2,3). During the last decade, additional oncogenic drivers have been identified in NSCLC, such as *ALK*, *ROS1*, *RET* and Neurotrophic Receptor Tyrosine Kinase *NTRK* gene rearrangements or *BRAF*, *KRAS*, *ERBB2* or *MET* mutations, that are associated with clinical benefit from specific targeted therapies (4). Current European guidelines recommend molecular testing in all patients with advanced lung

adenocarcinoma and in patients with squamous cell carcinoma who do not have significant tobacco exposure or were younger than 50 years old. In this context, the incorporation of massive parallel sequencing (NGS) has become the most cost efficient and appropriate approach to screen oncogenic actionable alterations in NSCLC (5). A reduction of population-level mortality from NSCLC in the United States was observed from 2013-2016 due to treatment advances and, particularly, to the incorporation of targeted therapies (6).

Current studies show that about 50-60% of patients with advanced NSCLC harbor potentially actionable driver alterations (7). Despite oncogenic addiction driven by those genomic alterations, resistance to targeted therapies will eventually emerge. Furthermore, clinical outcomes are variable in tumors driven by similar genomic aberrations (8).

The treatment paradigm for patients with advanced NSCLC who did not harbor actionable alterations has also substantially changed with the incorporation of immunotherapy as a novel treatment option (9). However, responses to immunotherapy are heterogeneous, and the percentage of long-term survivors is still rather low in lung cancer (10). Despite its limitations, PD-L1 expression is a well-established biomarker currently used in the clinic for predicting response to immunotherapy in NSCLC. Additional biomarkers, such as tumor mutational burden (TMB) or the presence of certain genomic alterations (i.e., *STK11*, *KEAP1*, *PTEN*, *TP53*), have been also correlated

with clinical benefit or resistance to immunotherapy. However, these biomarkers, on their own, have been unable to fully predict clinical response or better survival outcomes upon immunotherapy regimens alone or in combination with chemotherapy (11). In this context, the incorporation of new layers of information could contribute to better capture the complexity of this disease and, ultimately, help to improve NSCLC patients' selection and clinical management.

In this chapter, some facts and figures regarding lung cancer epidemiology are introduced, as well as a summarized explanation of the current treatment landscape, focusing on patient stratification strategies for clinical decisions and treatment guidance in NSCLC. Finally, the unmet needs of the current patient selection framework can be found in section **2.1.4**.

2.1.1 Non-small cell lung cancer epidemiology

Lung cancer (LC) is a major public health problem. According to the World Health Organization (WHO) and the International Agency for Cancer Research (IARC), LC was the second most frequently diagnosed cancer in 2020 worldwide, with more than 2,000,000 cases (11,4% of all cancer diagnosis, excluding non-melanoma skin cancer (NMSC)) (**Figure 1**) (12).

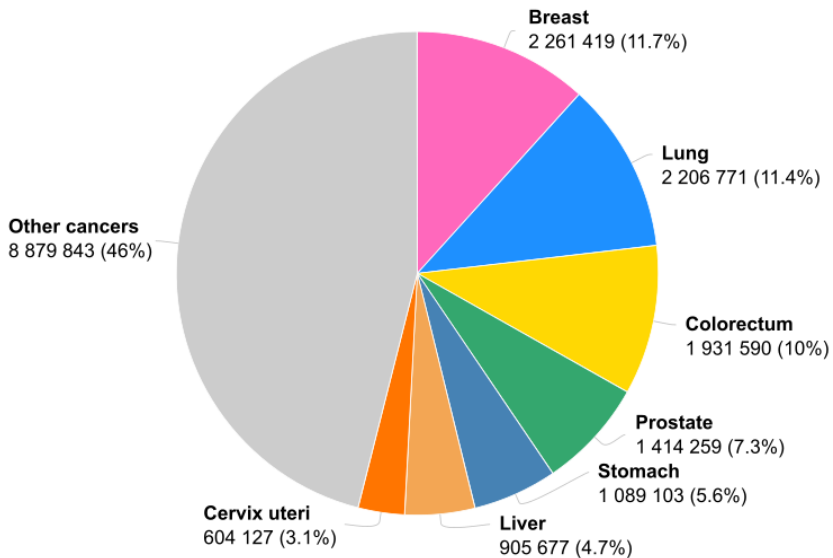


Figure 1. Estimated number of new cancer cases worldwide during 2020.

Pie chart representing the estimated percentage share of the most frequently diagnosed cancer types in the total number of cancer-related diagnoses made in 2020. Estimation was performed considering both sexes, all ages and excluding non-melanoma skin cancers. (Source: Adapted from (12)).

In Western countries, the five-year overall survival rates for patients with LC are strongly correlated with tumor stage at the time of diagnosis, ranging from 61.2% in patients with localized disease to 33.5% in patients with locally advanced disease, and down to 7.0% for patients with disseminated disease (13) (Figure 2).

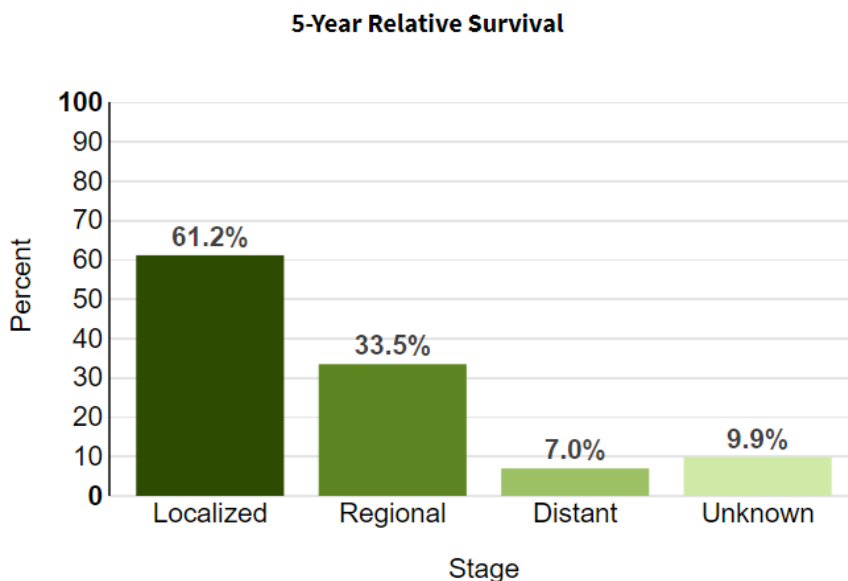


Figure 2. 5-year relative survival by stage at diagnosis.

5-year relative survival based on data from the National Cancer Institute Surveillance, Epidemiology and End Results (SEER) between 2012 and 2018. Survival rates were calculated considering both sexes and all races. (Source: Extracted from (13)).

At the time of diagnosis, the majority of patients with lung or bronchus cancer show regional or distant dissemination, which subsequently has a dramatic impact on the overall prognosis and life expectancy of the disease (13). Thus, the fact that these diagnoses occur late in the course of the disease, contributes to LC being the most frequent cause of cancer-related deaths worldwide, with almost 1.8 million annual deaths (12). Moreover, these numbers are expected to increase, reaching more than 3,500,000 diagnoses and almost 3,000,000 deaths globally by 2040 (12) (**Figure 3**).

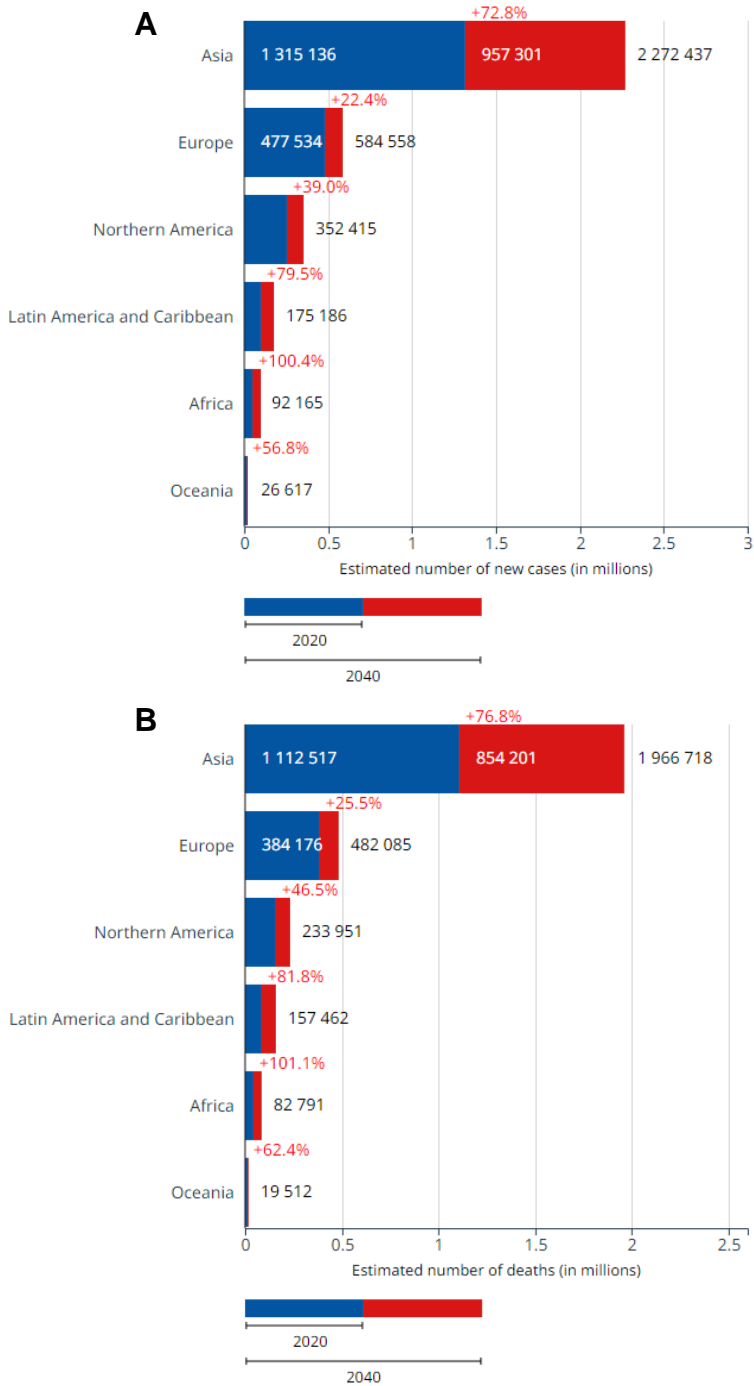


Figure 3. Estimated number of (A) incident cases and (B) deaths worldwide by 2040.

Stacked bar chart depicting expected LC incidence and mortality (in millions) stratified by continent. Estimation was performed considering both sexes and the whole age range [0-85+]. Blue and red bars represent figures in 2020 and the increment by 2040, respectively. In addition, the increase rate from 2020 to 2040 is highlighted in red for each continent category. (Source: Extracted from (12))

2.1.2 Towards an increasingly accurate non-small cell lung cancer classification: from histological to molecular subtyping

LC was traditionally classified into NSCLC and small cell lung cancer (SCLC), accounting for 85% and 15% of LC cases, respectively (14). The incorporation of molecular testing and personalized medicine into the clinical management of NSCLC led to further stratification of lung tumors beyond histological features. In this way, the 2015 WHO classification of lung tumors divided NSCLC into three major histological subtypes: lung adenocarcinoma (LUAD ~ 40% of NSCLC cases), lung squamous cell carcinoma (LUSC ~ 25-30 % of NSCLC cases) and large cell carcinoma (LCC ~ 10-15% of NSCLC cases). Another subtype is “not otherwise specified” (NOS) which has none of the specific characteristics of the aforementioned subtypes. This classification determines eligibility for further molecular testing and selection for certain therapeutic strategies (15).

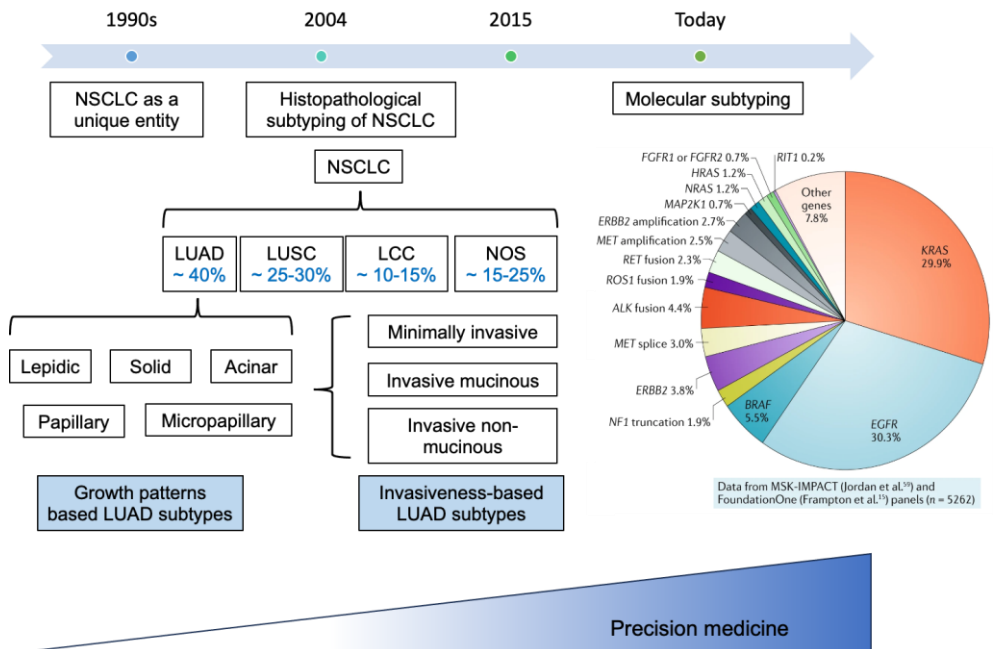


Figure 4. Non-small cell lung cancer subtyping evolution: from histological to molecular stratification.

Timeline highlights some of the most important advancements towards increasingly comprehensive non-small cell lung cancer stratification (Source: modified from (16))

Advances in DNA- and RNA-based high-throughput genomic technologies for molecular profiling have also improved the classification and clinical management of NSCLC, especially for advanced LUAD (17). Specifically, the identification of epidermal growth factor (*EGFR*) mutations (~14%) and anaplastic lymphoma kinase (*ALK*) rearrangements (~2-7%), as well as their association with response to selective tyrosine kinase inhibitors (TKIs) represented a paradigm shift in the treatment of LUAD and transformed the pathological classification (18,19). These findings were followed by the

discovery of additional actionable oncogenic alterations, including KRAS G12C (~14%), BRAF V600E (~ 2-3%) and *ERBB2* (~ 2%) mutations, *MET* amplification and exon 14 skipping mutations (~ 7%) and gene fusions involving *ROS1* (~ 1%), *RET* (~ 1%) and *NTRK* (< 1%) genes (20,21). Current European clinical guidelines recommend to screen for alterations in all the previously mentioned oncogenic drivers (4). In the case of advanced LUSC, the frequency of driver alterations is much lower and genomic testing is only recommended for patients younger than 50 years old or with low tobacco exposure (4).

Finally, an updated LC classification has been published as part of the 2021 WHO Classification of Thoracic Tumors (22). Briefly, the major features among this new edition include the even greater emphasis on genomic profiling than in the previous 2015 WHO classification, due to the development of high-throughput screening techniques and the discovery of new and more effective targeted therapies. A timeline representing the evolution of lung cancer classification from histological to genomic-based subtypes is depicted in **Figure 4**.

2.1.3 Immunotherapy in non-small cell lung cancer: the revolutionary weapon

During the past decades, many scientific advances have contributed to increase our understanding of the underlying

biology of NSCLC. Therefore, we assisted to a paradigm shift in the first-line treatment of advanced/metastatic NSCLC patients. For instance, this revolution started with the identification of targetable oncogenic alterations, as mentioned in section **2.1.2**. In this way, considering the currently approved range of targeted therapies, approximately one third of patients with advanced LUAD could be eligible for biomarker-directed therapies (23). However, there is still a considerable proportion of patients that cannot benefit from these targeted treatments. This is especially true for LUSC, where the frequency of targetable alterations is much lower than in LUAD (24). In this context, immunotherapy, and more specifically, immune-checkpoint inhibitors (ICIs) targeting PD-1/PD-L1 and CTLA-4, alone or in combination with chemotherapy, has become the standard of care in the frontline for patients with advanced/metastatic NSCLC without actionable driver alterations (25).

NSCLC tumors can activate the PD-1/PD-L1 molecular pathway through the adaptive immune resistance, where cancer cells change its phenotype in response to a cytotoxic or pro-inflammatory immune response, leading to immune evasion (26). This adaptive mechanism is promoted by the specific recognition of cancer cells by T cells, that secrete immune-activating cytokines. Cancer cells express PD-L1 to protect themselves from the T cell attack, whereas PD-1, its receptor, is preferentially expressed in T cells. PD-1 activation leads to the reduction on T cell proliferation, cytokine

production and T cytotoxic functions (27). In the end, these signals produce an immunosuppressive environment and contribute to tumor progression and development. Immune checkpoint inhibitors, such as anti-PD-1/PD-L1 antibodies, are able to overcome this inhibitory signal and can restore immune response, through the reactivation of intratumoral pre-existing T cells turned off by adaptive immune resistance. This T cell re-invigoration can ultimately lead to tumor destruction (28).

It has been less than ten years since the anti-PD-1 antibody nivolumab received FDA-approval for the treatment of patients with LC (29). This marked the beginning of a new era in the clinical management of NSCLC, especially for those patients not benefiting from targeted therapies. This approval came after the results from CheckMate-017 and CheckMate-057, two randomized clinical trials which demonstrated that nivolumab was able to improve median OS of platinum-resistant squamous NSCLC and non-squamous NSCLC, respectively, compared with docetaxel in the second-line setting (30,31). Notably, a combined analyses of data from both CheckMate-017 and CheckMate-057, showed that patients with nivolumab regimen had a 5-year OS rate of 13.6% compared with 2.6% for those patients receiving docetaxel (32). However, although PD-1 blockade revealed unprecedented long-term responses, they were observed only in a minority of patients with advanced NSCLC.

After nivolumab positive results for platinum-resistant NSCLC patients, immune checkpoint inhibitors were evaluated in the frontline setting and were compared to platinum-based chemotherapy which has been the standard of care for decades. Several studies demonstrated the superiority of anti-PD1 or anti-PD-L1 treatments over chemotherapy in the first-line setting (28). However, patient selection based on PD-L1 expression and molecular testing is crucial for treatment planning in the frontline, since not all patients will benefit from upfront immunotherapy alone. In this regard, there is still room for improvement, especially when selecting patients who might best benefit from these treatments and to define the most successful combinations for each individual patient.

2.1.4 Limitations of the current management framework of non-small cell lung cancer patients

Current clinical management of patients with NSCLC considers three main factors: clinicopathological characteristics (i.e., histology, tumor stage and location, age, performance status, comorbidities, organic function), presence of druggable oncogenic alterations (i.e., *EGFR*, *KRAS G12C*, *BRAF*, *ALK*, *ROS1*, *RET*, *NTRK*, *MET*) and PD-L1 status (5,33).

Although the introduction of molecular testing within the treatment decision framework has revolutionized and improved NSCLC patient stratification, this approach has some intrinsic

limitations that should be addressed. For instance, patients with the similar genotype often follow different disease courses for reasons that remain unclear (34). Also, all patients receiving targeted therapies will eventually develop treatment resistance for reasons that genomic profiling cannot always explain (i.e., activation of pathways acting downstream or parallel to the inhibited target, interactions with the tumor immune microenvironment (TME)) (35). Finally, more importantly, there is still a large group of NSCLC patients that do not harbor targetable genomic alterations and that cannot, therefore, benefit from available targeted therapies.

Immunotherapy alone or combined with chemotherapy is currently the standard of care for advanced NSCLC without targetable oncogenic alterations (25). However, the proportion of long-term survivors remains low (~20%), probably due to insufficient patient selection and lack of understanding of the complex interplay between tumor cells and tumor microenvironment. In this context, currently used individual biomarkers, such as PD-L1 IHC expression or tumor mutational burden (TMB), are not able to accurately predict immune response clinical benefit for a large proportion of patients (11).

In this context, the implementation of new methodologies beyond genomic testing, such as those based on global gene expression, could be very useful to further understand disease

complexity and, in the end, to deliver more precise and effective treatments to NSCLC patients.

2.2 Transcriptional profiling as a tool for tackling complexity and molecular heterogeneity in lung adenocarcinoma and lung squamous cell carcinoma

In **section 2.1**, NSCLC has been introduced as a complex and molecularly heterogeneous disease. It has been shown that the discovery of oncogenic genomic alterations in landmark genes (i.e., *EGFR*, *KRAS*, *BRAF*, *ERBB2*, *ALK*, *ROS1*, *RET*, *NTRK*, *MET*) that confer sensitivity to specific therapeutic strategies has improved both life quality and expectancy of patients with advanced NSCLC. However, many patients with advanced NSCLC do not harbor actionable genomic alterations (50%) and are treated with concurrent or sequential chemotherapy and immunotherapy, thus not receiving personalized treatments.

In this context, the study of the transcriptome could be a relevant tool for improved patient stratification that could enable the transition to a more personalized approach. Underpinned by the idea that gene expression analysis provides great understanding of cellular processes and tumor biology, transcriptional profiling has been extensively used to dissect tumor heterogeneity and define clinically relevant

subtypes within a cancer entity. Examples of these include breast cancer intrinsic subtypes or colorectal cancer consensus subtypes (36,37). In lung cancer, initial studies demonstrated the ability of transcriptional profiling to recapitulate histological subtypes (38,39). Seminal studies were able to further classify LUAD into molecularly different subtypes based on differential gene expression (40–42). Although those studies included a majority of LUAD tumors that could have hampered the identification of transcriptional groups in other histologies, later studies focused on one histology, generally LUAD or LUSC, due to their higher incidence (43).

In this chapter, the current state-of-the-art of transcriptional-based classifications in the context of NSCLC major subtypes (i.e., LUAD and LUSC) will be revised, as well as the clinical relevance of these classifications.

2.2.1 Lung adenocarcinoma transcriptional subtypes

Given the greater transcriptional heterogeneity previously observed for LUAD and its higher prevalence, many studies have attempted to generate a more refined classification of these tumors (**Figure 5**) (38,39,44–60). Overall, tumors were classified according to the gene expression levels of the most highly variable genes between samples. Diverse LUAD intrinsic transcriptional-based subtypes were identified across distinct studies. All of them observed a group of well-

differentiated LUAD tumors with higher levels of pneumocyte-related markers and associated with better survival outcomes. Also, they identified a subset of poorly differentiated LUAD tumors with high expression of proliferation-related genes and associated with poor survival. However, despite these commonalities, each of these studies also identified unique subtypes which weren't replicated in other works. In addition, the genes defining phenotypically comparable subtype specific signatures differed between studies, which reduces reproducibility and robustness.

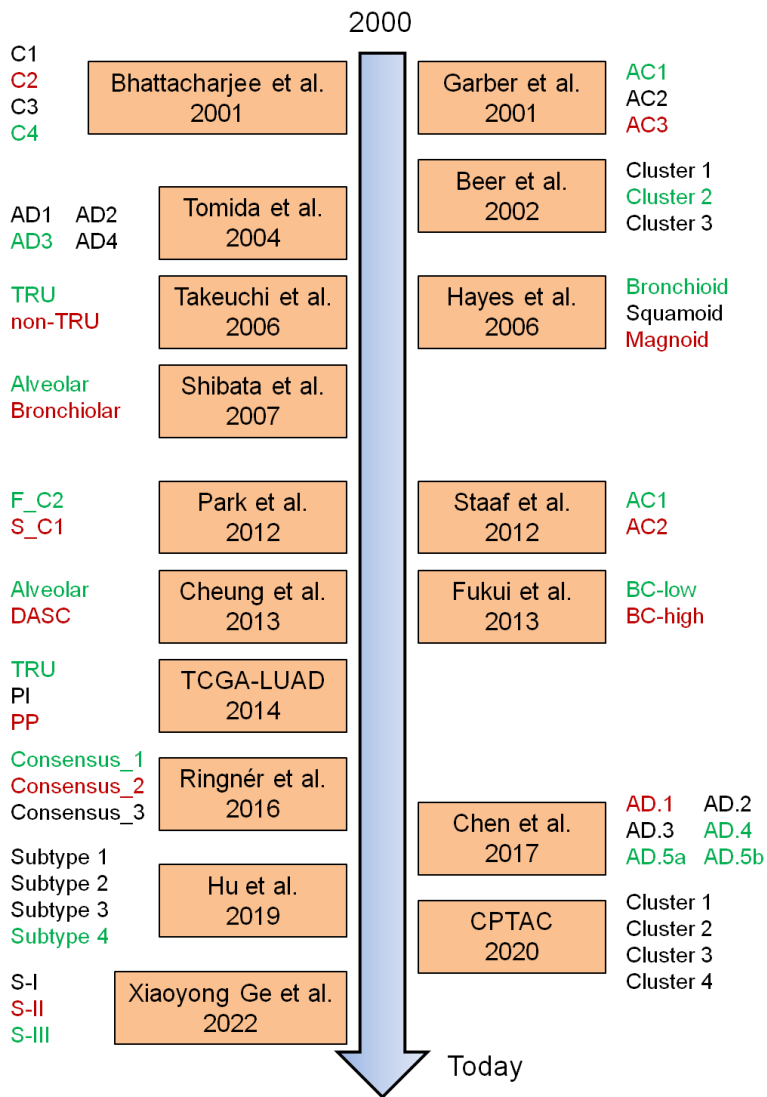


Figure 5. Transcriptional-based lung adenocarcinoma molecular subtypes.

Timeline highlights transcriptional-based lung adenocarcinoma subtypes described by different studies since year 2000. Identified subtypes are depicted next to the corresponding study, represented by an orange box. Red color represents subtypes associated with proliferative tumors and unfavorable prognosis, whereas green color represents subtypes associated with less proliferative tumors and favorable prognosis.

To date, the most accepted transcriptional-based classification in LUAD was described back in 2006 by Hayes et al. (47). This classification was further validated by The Cancer Genome Atlas (TCGA) Consortium, and it is also considered as the TCGA consensus classification (55). In summary, Hayes et al. described three LUAD subtypes: **bronchioid, squamoid, and magnoid**. These subtypes differed in terms of survival outcomes and molecular characteristics. In this way, **bronchioid subtype** was associated with better prognosis compared with the other two groups and had higher expression of cisplatin-resistance genes. **Squamoid tumors** were characterized by overexpression of angiogenesis-related genes and WNT pathway. Finally, **magnoid subtype** showed higher expression of genes involved in inflammatory processes, metabolism, cytoskeleton remodeling and proliferation. Wilkerson et al. investigated whether these intrinsic subtypes were correlated with particular genomic alterations (52). They found that the **bronchioid subtype** was enriched for *EGFR* alterations, while the magnoid subtype had higher frequency of *KRAS* and *TP53* mutations, higher levels of genome instability, copy number alterations rate, DNA hypermethylation and TMB compared with the other subtypes.

In 2014, the TCGA Consortium classified 230 LUAD tumors according to the Hayes et al. classification (47,55). LUAD subtypes were renamed to better represent the histological, morphological, and molecular characteristics of the three different subtypes: **terminal respiratory unit (TRU,**

bronchioid), proximal inflammatory (PI, squamoid), and proximal proliferative (PP, magnoid). Comprehensive molecular characterization of the subtypes corroborated the association of the **TRU subtype** with the presence of *EGFR* alterations and also *ALK* rearrangements. Patients belonging to **PI intrinsic subtype** were enriched for *KRAS* and *STK11* mutations, whereas in the **PP subtype** concurrent mutations in *NF1* and *TP53* genes were observed.

Hu et al. elaborated a framework for a transcriptional-based stratification of LUAD tumors using the new TCGA-LUAD data and defined 4 subtypes (58). A pathway enrichment analysis of the most representative genes revealed that groups 1 and 2 showed higher activity of immune-related pathways, whereas subtypes 3 and 4 overexpressed pathways involved in cell proliferation and extracellular matrix organization, respectively. Concerning oncogenic alterations, *TP53* mutations were more frequently found in tumors belonging to subtypes 1 and 2, while group 4 were enriched for *EGFR* mutations. In addition, multivariate analyses demonstrated that these groups had independent prognostic value.

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) project conducted a study in which they combined molecular data from different sources (i.e., gene expression, genomics, proteomics, phosphoproteomics) of 110 LUAD tumors and identified four intrinsic molecular subtypes (59). Moreover, integration of the different molecular information layers,

especially protein phosphorylation and acetylation modifications revealed potential tumor-specific markers and druggable proteins.

Also following a multi-omics approach, combining transcriptional profiling with other molecular data, Chen et al. reported another LUAD classification consisting of six LUAD subtypes: AD1-AD4, AD5a, and AD5b (57). In agreement with previous reports, those subtypes associated with lower differentiation levels demonstrated worse survival rates. Additionally, genes representing each identified subtype were associated with specific transcriptional programs and biological processes.

In another study, Xiaoyong Ge et al. developed a LUAD classification based on the expression of treatment associated genes extracted from cancer cell lines pharmacogenomics data on the TCGA-LUAD dataset (60). This analysis resulted in the identification of three LUAD subtypes: S-I, S-II, and S-III, which were associated with different clinical features and previously described TCGA consensus subtypes (55). S-I displayed overexpression of inflammation-related genes and was associated with proliferation and immune evasion. S-II showed overexpression of cell cycle-related genes and was associated with higher mutation burden. Finally, S-III demonstrated higher expression of metabolic signatures and its development was associated with methylation processes. Interestingly, this study tried to assign potential specific

treatments for these subtypes using the CMap database (61). In this way, immune checkpoint blockers (ICB), doxorubicin, tipifarnib, AZ628, and AZD6244 were found to be potentially effective for S-I tumors; cisplatin, camptothecin, roscovitine, and A.443654 seemed to work for S-II subtype; and S-III tumors could be potentially sensitive to docetaxel, paclitaxel, vinorelbine, and BIBW2992.

Overall, all these classifications highlight the inherent heterogeneity of LUAD, which cannot be considered a unique histological entity. Therefore, a deeper understanding of this diversity and further stratification beyond histological and genomic features is needed as it could help to improve the clinical management of LUAD.

2.2.2 Lung squamous cell carcinoma transcriptional subtypes

Although most studies regarding transcriptional subtypes definition focused on LUAD due to its higher prevalence, some studies also reported transcriptional-based subtypes for LUSC (**Figure 6**). In this way, seminal studies identified a transcriptional-based classification of LUSC consisting of two main groups associated with different survival outcomes and differentiation grades (46,62,63).

In 2010, the group of Wilkerson et al. reported four LUSC intrinsic subtypes based on differential gene expression: **primitive, secretory, basal, and classical** (64). In the same way as in LUAD, these subtypes, were further characterized

by TCGA and, therefore, this classification is commonly known as the TCGA-LUSC consensus classification (65). Overall, these subtypes displayed different survival rates, differentiation grades and a specific transcriptional footprint. **Primitive tumors** showed higher expression of cell cycle-related genes and those patients had the worst prognosis. **Classical subtype** had overexpression of genes involved in xenobiotic metabolism. **Secretory-like tumors** were associated with immune system-related signatures and pneumocyte type II markers expression. **Basal subtype** displayed higher expression levels of cell adhesion and basement membrane function associated genes. In a subsequent study by Brambilla et al., an additional basaloid-like subtype was also identified (66). This subtype demonstrated a high correlation with Wilkerson et al.'s primitive subtype but with an even more strong association with poor survival rates.

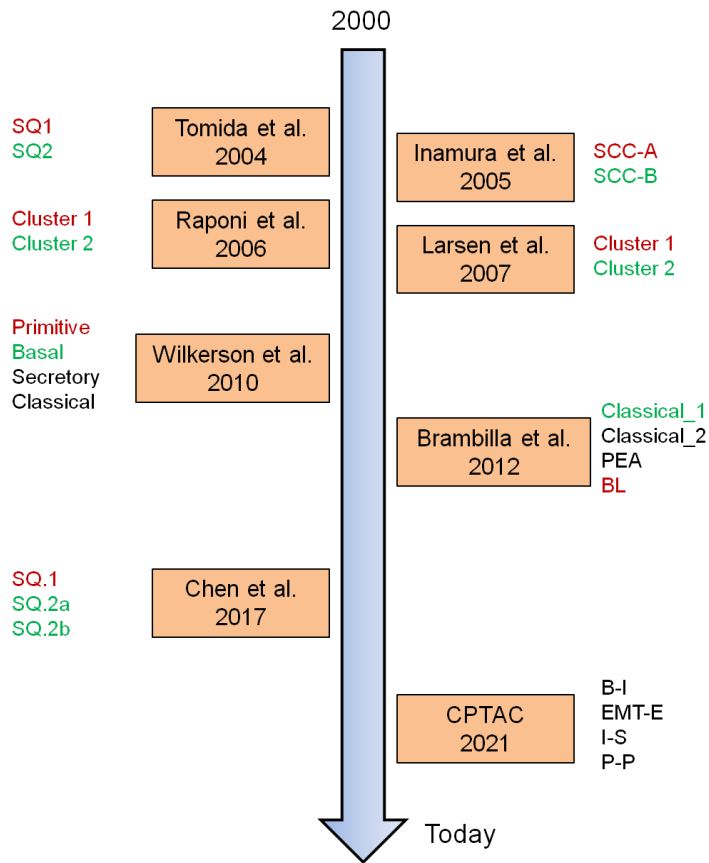


Figure 6. Transcriptional-based lung squamous cell carcinoma molecular subtypes.

Timeline highlights transcriptional-based lung squamous cell carcinoma subtypes described by different studies since year 2000. Identified subtypes are depicted next to the corresponding study, represented by an orange box. Red color represents subtypes associated with proliferative tumors and unfavorable prognosis, whereas green color represents subtypes associated with less proliferative tumors and favorable prognosis.

As previously mentioned, following Wilkerson et al.'s classification proposal, TCGA conducted a comprehensive characterization of the subtypes in almost 200 LUSC samples, especially regarding the association with genomic features and differential methylation profiles (65). For instance, **classical**

subtype was enriched for *KEAP1*, *NEF2L2* and *PTEN* mutations, strong hypermethylation and overall genomic instability, compared to other LUSC classes. On the other hand, **primitive tumors** were enriched in *RB1* and *PTEN* DNA alterations, while **basal subtype** displayed higher *NF1* mutation frequency.

Finally, the combination of multiple sources of molecular data, including transcriptional profiling, improved our understanding the complex biology of LUSC tumors. For instance, the CPTAC study reported a multi-omics classification into five groups based on DNA, RNA, protein, phosphoprotein, and acetylprotein information in 108 LUSC: **basal-inclusive (B-I)**, **epithelial-mesenchymal transition-enriched (EMT-E)**, **classical**, **inflamed-secretory (I-S)** and **proliferative-primitive (P-P)** (67). **B-I subtype** comprised basaloid-like tumors with high expression of metabolic, immune, and estrogen response-related genes. **EMT-E subtype** was characterized by the expression of EMT, angiogenesis and myogenesis signatures and included tumors with myxoid histology and fibroblast infiltration. **Classical subtype** was associated with higher mutation frequency of *KEAP1*, *CUL3* and *NFEL2L*, copy number amplifications in *SOX2* and *TP63* genes, hypermethylation, and with the previously described TCGA-LUSC classical subtype. Moreover, **classical tumors** displayed higher activity levels of oxidative phosphorylation and proliferation signaling pathways, as well as a low activation of immune-system related signatures. **I-S subtype** was

strongly associated with the secretory TCGA-LUSC subtype and demonstrated upregulation of immune-related genes. Finally, **P-P subtype** showed upregulation of cell cycle-related genes and downregulation of immune system-related pathways, as well as patterns of DNA hypomethylation. Furthermore, this study integrated data from pharmacogenomic studies and drug databases to identify potential treatments for LUSC tumors. However, although some associations were established, therapeutic vulnerability identification efforts were not subtype-centered. Finally, Chen et al. also reported a LUSC classification in three groups using a multi-omics approach: SQ.1, SQ.2a and SQ.2b. SQ.2a and SQ.2b displayed similar expression patterns but different methylation profiles (57). Also, as reported by CPTAC, Chen et al. subtypes were associated with *SOX2* and *TP63* genes and targets. In terms of prognosis, SQ.1 was associated with shorter overall survival. Moreover, associations between these subtypes and the more accepted TCGA-LUSC classification, revealed a correlation between basaloid and secretory subtypes with SQ.1, classical subtype with SQ.2a and classical and primitive subtypes with SQ.2b.

Analogously to LUAD, the identification of different transcriptional subtypes in LUSC reflects the molecular heterogeneity that exists within this often-considered homogeneous disease.

2.2.3 Clinical relevance of lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based classifications

Despite all the efforts mentioned in sections **2.2.1** and **2.2.2** to establish a transcriptional-based classification of LUAD and LUSC, the reality is that these classifications have not reached clinical practice. One of the main concerns is their low reproducibility and the limited concordance across different studies. In this way, transcriptional subtypes derived from these studies, were generally conducted in a limited number of samples, were based on differential gene expression analyses that provide large lists of genes that, after passing significance threshold, are considered to represent each of the defined groups. Individual gene expression measures are subjected to multiple sources of technical and biological variability, such as the use of different gene expression platforms, and this could partially explain the lack of reproducibility. Moreover, the interpretation of classifications based on large sets of genes is complex, and these types of tests are also difficult to be implemented in the clinic (68), where the use of fresh tissue is scarce. For instance, little overlap was found between the gene expression signatures derived from some of these studies, even for allegedly correlated subtypes (**Figure 7**) (69).

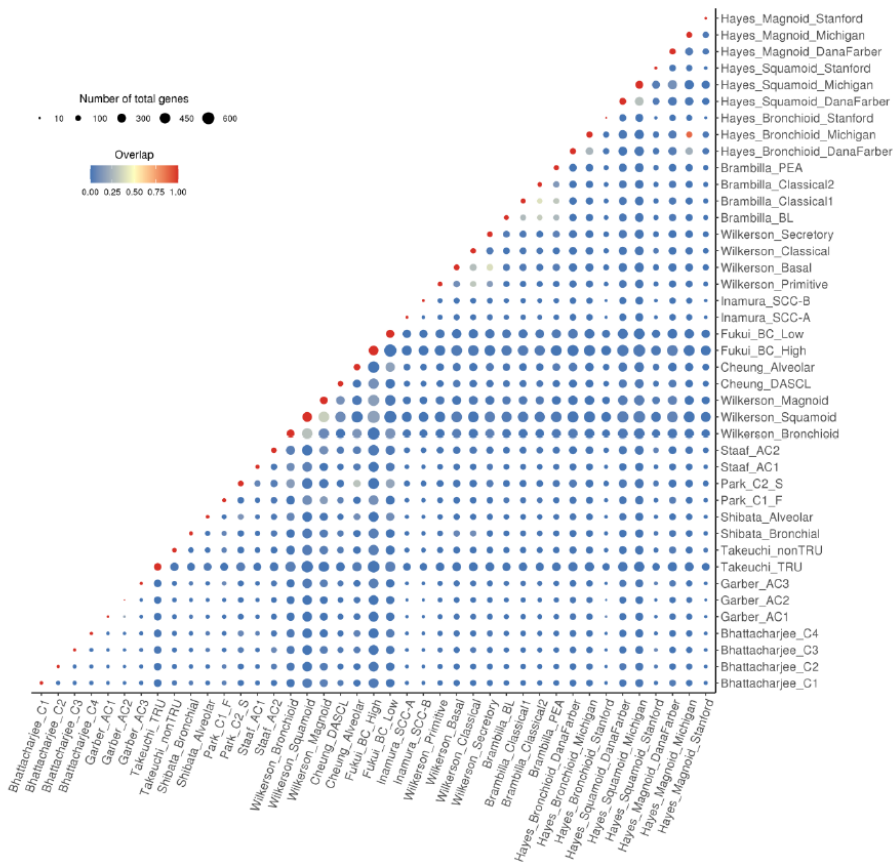


Figure 7. NSCLC transcriptional subtypes gene expression signatures overlap.

(Source: Extracted from (69)). Size of the dots represents the number of genes included in each subtype signature and color indicates the degree of overlap between the signatures defining specific subtypes in the different studies. Blue colors depict lower overlap degree, whereas red colors indicate higher concordance between a specific pair of signatures.

In addition to the high complexity of the analysis and interpretation of transcriptomics data, the emergence of genomic profiling and the identification of actionable oncogenic drivers reduced the interest on gene expression profiling (3). However, the most important limitation of all the transcriptomic

classifications of lung cancer is the lack of therapeutic impact of the intrinsic subtypes. This leads to most patients who do not harbor actionable mutations being equally treated with chemotherapy and immunotherapy. Moreover, prospective validation of those subtypes and their impact on specific therapeutic vulnerabilities would be needed before being incorporated into NSCLC routine clinical management. Finally, an additional limitation is that most studies are retrospective and utilized microarray and RNA sequencing (RNA-Seq) techniques, which normally rely on fresh-frozen (FF) tissue samples. For this reason, most studies are enriched in early or locally advanced tumors, in which surgical resection can be conducted. Also, access to FF samples is difficult in the clinical practice, which also limits the implementation of transcriptional subtypes into other type of specimens like formalin-fixed paraffin embedded (FFPE) samples. Finally, access to systemic therapies is limited in earlier clinical settings and therefore unravelling a correlation between intrinsic subtypes and specific treatments was indeed very unlikely.

2.3 Potential applications of transcriptional profiling in the clinical setting: adding another layer of information

In **section 2.2** we have seen that, LUAD and LUSC are not unique diseases but constitute multiple molecular entities, each with a unique transcriptional footprint. Current clinical management of NSCLC, based on histological features and specific genomic alterations, only captures partly the inherent heterogeneity of this disease. This is especially evident when variable outcomes and treatment responses are observed between patients with identical histology or genotype. In this context, the implementation of additional layers of information, such as transcriptional profiling, could be helpful to capture this heterogeneity and, in the end, to deliver more specific and effective treatments to patients with NSCLC.

This chapter focuses on how gene expression technologies can help to explain and tie up some of the loose ends that still exist. For instance, its use in the context of patients without actionable alterations, drug resistance mechanisms and patient stratification for immunotherapy and targeted therapies will be discussed throughout the different sections.

2.3.1 Patients lacking actionable driver alterations: opening a window for new therapeutic strategies

Comprehensive genomic profiling, including not only the identification of single nucleotide variants (SNV) but also gene fusions and splice variants, has transformed the clinical management of patients with non-squamous NSCLC and is currently needed to guide treatment decisions (70). The identification of novel targets as well as the emergence of genomic alterations during tumor progression and subsequent drug development, is contributing to the enlargement of diagnostic gene panels that, consequently, are becoming more comprehensive. Most complete currently available panels offer the possibility to analyze DNA and RNA alterations, and more recently, tumor mutation burden, homologous recombination deficiency or microsatellite instability (71). However, these panels still cover a specific and reduced set of genes. Thus, the implementation of whole transcriptome RNA-Seq could potentially allow for the detection of yet non-described fusions or other actionable molecular events not accounted by the panels in patients classified as driver-negative by these targeted DNA/RNA-Seq techniques (72).

In addition to sensitivity issues due to the limit of detection for certain alterations, around 50% of patients do not harbor actionable alterations and therefore cannot benefit from targeted therapies (7). This is an important issue for patients

with LUSC, in which actionable alterations are anecdotal, and for a non-negligible number of patients with LUAD. For this reason, there is a need to find new therapeutic vulnerabilities in NSCLC, otherwise uniformly treated with chemoimmunotherapy. In this way, transcriptional profiling has demonstrated to be very useful to guide treatment decisions, especially in those cases with advanced tumors that have recurred upon treatment. For instance, two recent studies concluded that incorporating information from gene expression technologies upon progression would increase the rate of patients benefiting from targeted therapies, when compared to using only DNA (73,74).

Furthermore, in a recent prospective study published in 2022 a combination of whole genome and transcriptome analyses (WGTA) was used to align treatments to 570 patients with metastatic/advanced tumors, from which 67 (12%) were lung cancer patients, and that had received prior therapy (75). Integration of these data allowed the identification of 514 alterations and 248 associated therapies. Actionable targets were identified in 83% of tested patients (475/570) and 37% (209/570) received corresponding targeted therapy. In this study, transcriptomics data proved to be very revealing, being the most common data source contributing to WGTA-informed treatments (67%, 168/248). Moreover, 25% (63/248) of treatments were based exclusively on transcriptional information. Meanwhile, genomic mutations contributed to the identification of only 34% (85/248) of WGTA-informed

treatments, with 14% being discovered through mutation data alone (**Figure 8**).

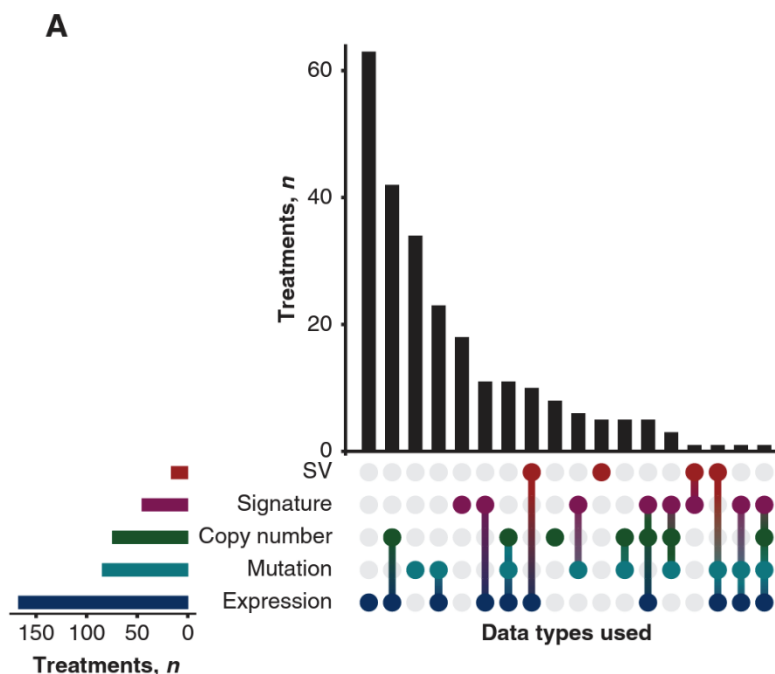


Figure 8. UpSet plot describing the number of treatments identified using different data types, alone (indicated by filled circles) or in combination (indicated by lines connecting different data types represented by filled circles).

(Source: Extracted from (75))

Results regarding final performance were impressive, with almost 50% of delivered treatments resulting in clinical benefit. It is important to highlight that in the above-mentioned studies, they use transcriptomics for the detection of specific genomic alterations, such as fusion events or splice variants, that are normally detected using RNA-Seq techniques. Thus, in this

case, the treatment is guided by the presence/absence of these alterations rather than by the classification of tumors based on their transcriptional profiling. This second approach would again be particularly interesting for those patients without specific actionable alterations who, in the case of advanced NSCLC, are normally homogeneously treated with chemotherapy and/or immunotherapy regimens. However, although potentially informative for treatment guidance, patient stratification based on transcriptional signatures, is still not a common practice in the clinical setting or clinical trials, as it requires more complex analysis and interpretation of the results.

Therefore, although challenging, the combination of genomics and transcriptomics data should be seriously considered for more precise patient stratification in the context of NSCLC clinical management. Moreover, although all the previously mentioned studies were conducted mostly under the setting of progressive disease, this approach might prove very informative in order to choose the appropriate first-line treatment regimen.

2.3.2 Anticipate and overcome potential treatment resistance mechanisms

Targeted therapies have revolutionized the treatment landscape in NSCLC and have led to an unprecedented improvement in patients' life expectancy. However, treatment resistance will eventually emerge, leaving patients with very

few treatment alternatives and poor overall prognosis once the disease has recurred. Resistance mechanisms can be on-target, when they are driven by the acquisition of genomic alterations that hamper the inhibition of the target by the drug, and off-target, when there is an activation of downstream or parallel pathways that overcome the blockade of the oncogenic protein (76) (**Figure 9**).

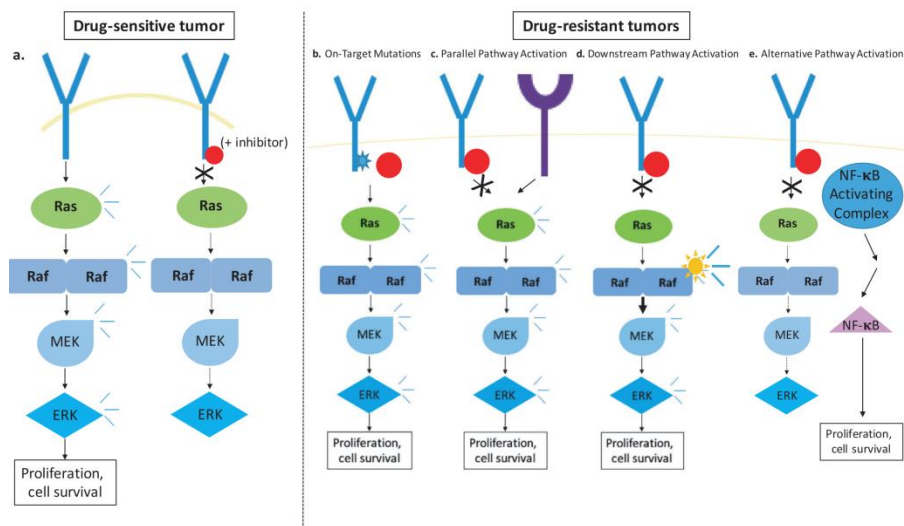


Figure 9. Mechanisms of resistance to targeted therapies.

A) Diagram representing the mechanism of action of Ras inhibitors in a drug-sensitive tumor. **B)** Diagram depicting different resistant mechanisms for Ras inhibitors in drug-resistant tumors (e.g. on-target mutations, parallel pathways activation, downstream pathway activation, alternative pathway activation) (Source: Extracted from (76)).

Current strategies for understanding the basis of tumor resistance to therapy include the development of preclinical models of tumor resistance (e.g., cell lines, mouse models, patient-derived organoids, etc.) and retrospective analyses of tumor samples obtained at progression. The development of novel sequencing techniques and bioinformatics approaches

led to the identification of different biomarkers that helps us understand how specific tumor molecular characteristics impact on the response to treatment. In this context, transcriptional profiling could be very useful to tackle off-target resistance. Indeed, many of these mechanisms may be originated at non-genomic levels, such as epigenetic modifications, tumor-TME interactions or histological transformations, but may have an impact on the transcriptome behavior (77,78).

EGFR-TKIs are widely used in the context of *EGFR* mutated NSCLC, and they have proven clinical benefit in this group of patients. Thus, overcoming and anticipating resistance is clinically relevant (79). In this way, preclinical studies have shown many potential off-target resistance mechanisms to EGFR-TKIs, including PI3K/AKT/mTOR or EMT pathways upregulation (80,81). However, the main limitation of these preclinical studies is that they were mainly focused in assessing the deregulation of one or two proteins involved in these pathways by targeted sequencing or molecular biology techniques. Unfortunately, these targeted approaches are less likely to unveil the mechanisms that lead to a specific pathway aberrant activation than untargeted methods which consider the whole transcriptome/proteome. Moreover, untargeted methods would also allow to assess whether other relevant pathways are being deregulated concurrently with the one under study. Therefore, the incorporation of whole transcriptome techniques might be crucial for the identification

of yet not described mechanisms of resistance that can be translated into new strategies to overcome tumor progression upon treatment. In this context, there are already some studies that have tried to unravel the mechanisms that lead to EGFR sensitivity or resistance by using microarray/RNA-Seq techniques covering the whole transcriptome (82–91). Briefly, these studies managed to generate gene expression signatures potentially able to predict EGFR-TKI efficacy. Nevertheless, the application of these signatures to the clinical setting is subjected to validation in prospective clinical trials, with matched gene expression and response rate information in NSCLC patients.

2.3.3 Patient selection for immunotherapy treatment regimens

Immunotherapy alone or in combination with chemotherapy is the standard of care in advanced/metastatic NSCLC without actionable genomic alterations (25). However, response rates to immunotherapy in NSCLC are around 20-40% and this is likely due to insufficient patient stratification and lack of sufficiently reliable predictive biomarkers (11,60). In this context, whole transcriptome sequencing could provide comprehensive information at the level of individual patients and could also help identify additional immunotherapy biomarkers for reliable and robust patient stratification in the clinical setting. In fact, the use of gene expression signatures to predict response to immunotherapy is being studied,

especially since transcriptional profiling data allows for the identification and characterization of the immune cell populations infiltrating the tumors. In this way, Tamborero et al. conducted a comprehensive classification of more than 9,000 tumor cases from 29 solid cancer types based on the expression of 16 gene signatures that represent 16 distinct immune cell populations (**Figure 10**) (92). Briefly, the six identified immunophenotypes ranged from very low immune infiltration, with low cytotoxic populations abundance, to high presence of almost all immune cell types, including cytotoxic cells. Furthermore, in this study, Tamborero et al., assessed the association of these immunophenotypes with additional gene expression and genomic features.

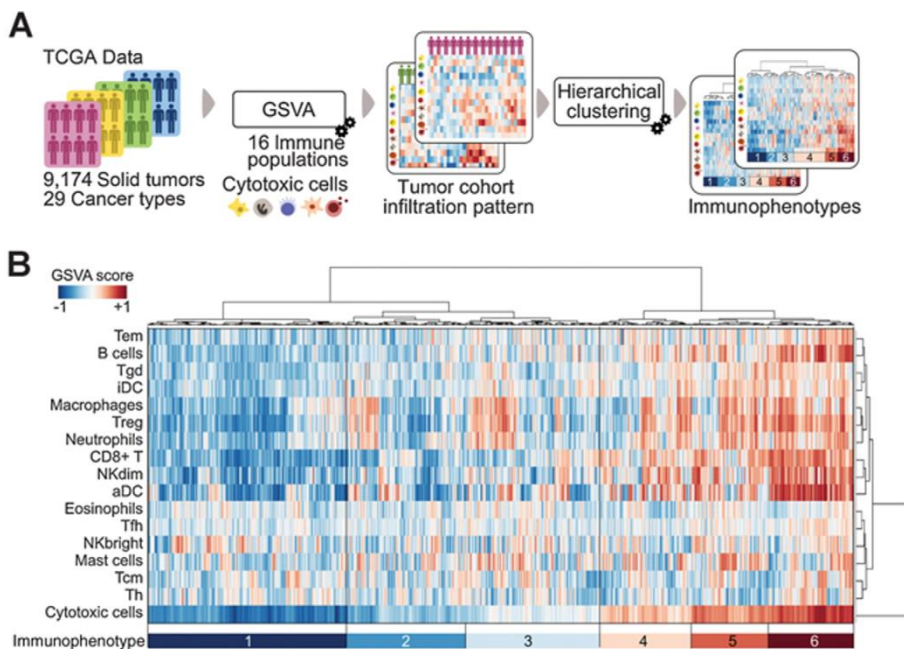


Figure 10. TCGA tumor immunophenotypes.

A) Methodology used for immunophenotypes definition in the different TCGA tumor cohorts. **B)** Heatmap representing the different immunophenotypes (1-6) based on the expression of 16 immune cell populations signatures. Red and blue colors indicate higher or lower infiltration levels of each cell population in each sample, respectively (Source: Extracted from (92)).

These biomarkers may help to understand how different patient subpopulations respond to immunotherapies, as well as to provide new evidence for the development of new strategies to boost immune response and overcome immunosuppressive scenarios. However, this study does not consider the impact of immunotherapy since most tumors from TCGA were surgically resected and patients did not receive any immunotherapy treatment.

A recent study focused on multi-gene signatures as predictors of immunotherapy response, demonstrated that the integration of multiple gene expression levels constituted a robust indicator of cytotoxic T cells infiltration in different solid tumors, including NSCLC (93). The IFN- γ (6-gene signature) and T-cell inflamed (18-gene signature) signatures derived from this study displayed better performances than those observed for PD-L1 immunohistochemistry in predicting response to pembrolizumab in PD-L1 unselected patients (**Figure 11**).

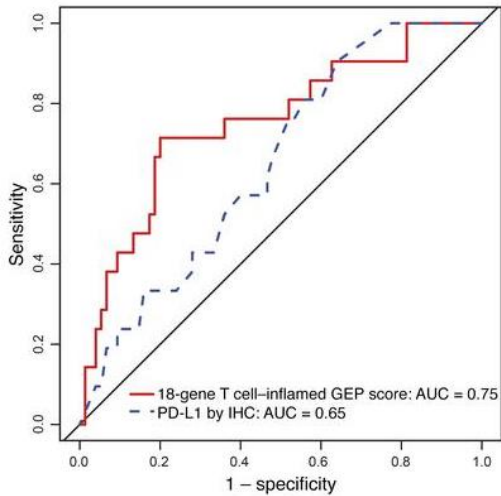


Figure 11. Evaluation of the predictive usefulness of the 18-gene T cell-inflamed GEP compared to PD-L1 IHC in predicting response to pembrolizumab in a PD-L1-unselected cohort of 96 patients with HNSCC from KEYNOTE-012

(Source: Extracted from (93)).

Hwang et al. conducted an immune profile targeted gene expression panel on previously untreated tumor samples to identify biomarkers that could potentially have an impact on immunotherapy response (94). After immunotherapy treatment, patients were divided into two different groups based on whether they had achieved a durable clinical response. Outcomes from gene expression analysis in these patients revealed that signatures associated with M1 macrophages and T cell infiltration were better predictors of response durability than currently used PD-L1 protein expression or TMB (**Figure 12**). However, these conclusions need to be validated in prospective clinical trials evaluating immunotherapy regimens.

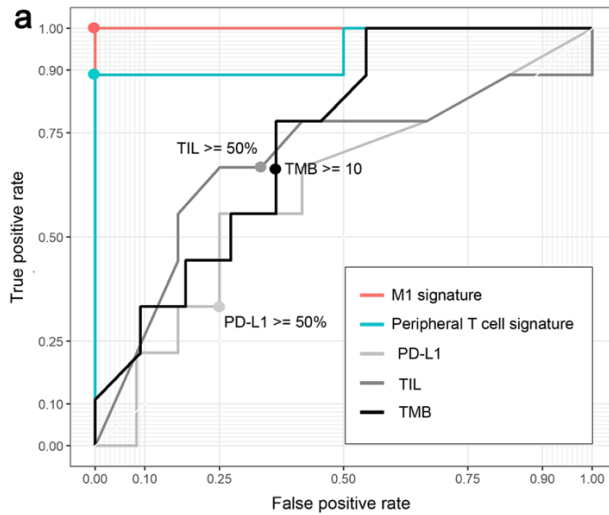


Figure 12. Predictive abilities of M1 signature, peripheral T cell signature, PD-L1 expression, tumor infiltrating lymphocytes (TIL), and tumor mutation burden (TMB)

(Source: Extracted from (94)).

Overall, these studies demonstrated the utility of multi-gene expression signatures over single agent predictors when evaluating patient's potential immunotherapy response. In this way, further datasets with available information on transcriptional profiling, treatment information and clinical response are needed for reliable biomarker candidates' identification. More importantly, future prospective clinical trials should start integrating these signatures as tools for improving patient selection for immunotherapy, so that they can be translated into the clinical practice in the future.

2.3.4 Patients harboring targetable genomic alterations: do all patients follow the same course upon treatment?

The commitment to precision medicine and the development of specific treatments targeting oncogenic pathways has transformed the current clinical management of NSCLC. The number of actionable genomic alterations to be tested in NSCLC have grown considerably (e.g., *KRAS*, *EGFR*, *ALK*, *ROS1*, *MET*, *RET*, *ERBB2*, *NTRK* and *BRAF*) based on the promising results observed in clinical trials evaluating specific inhibitors (95).

Genomic actionable alterations are found in about 17% - 80% of NSCLC, depending on the genetic ancestry, sex, smoking history, and histology (96). The identification of these mutations enables the use of specific treatments that have improved patients' survival outcomes and reduced undesirable adverse effects compared to standard chemotherapy regimens (97). Nevertheless, response rates vary between patients with the same oncogenic alteration and receiving the same treatment. Intratumoral heterogeneity and activation of additional molecular pathways beyond the one being targeted could partially explain these differences in response and duration of clinical benefit to targeted therapies (98).

Transcriptional profiling could be a good ally for further patient stratification in the context of patients harboring actionable genomic alterations and predict clinical benefit from targeted

therapies in this subgroup of patients. In this context, there are some studies that attempted to correlate NSCLC transcriptional subtypes with the frequency of actionable genomic alterations, mostly in the case of LUAD (48,49,52,55,57–59,65,67,99–101). These studies demonstrated that transcriptional stratification further refined genomic-based classification. In this sense, only two studies found specific transcriptional footprints for *EGFR* and *ALK* altered tumors, suggesting that the presence of these mutations may confer tumors a very specific transcriptional landscape (99,101). However, for most of the studies there was transcriptional heterogeneity in terms of gene expression for tumors harboring the same genomic alteration. For instance, *EGFR* mutated tumors were present across different transcriptional subtypes, although they were generally enriched in one of the identified groups. In these studies, tumors with different actionable alterations (i.e., *EGFR*, *KRAS*) were classified within the same transcriptional subtype, suggesting that different genomic alterations may lead to similar transcriptional landscapes and thus sensitivity to analogous drug targets. *EGFR*-driven tumors are considered a homogeneous population and are homogeneously treated, but variable drug responses and time to progression are observed in the clinic. Further stratification on these molecularly selected population might be useful for improving patients' clinical management.

Transcriptional heterogeneity was even more evident in tumors harboring *KRAS* mutations. *KRAS* represents the most mutated oncogene in non-squamous NSCLC, accounting for approximately 25% of LUAD in Western countries (21). Although *KRAS* had the reputation of an “undruggable” target due to its structural complexity, specific *KRAS* inhibitors for the G12C variant, such as sotorasib or adagrasib, represent a promising therapeutic option for patients with *KRAS*-G12C driven NSCLC (21). Given the great molecular heterogeneity observed for *KRAS*, it is reasonable to expect wide-ranging outcomes of those targeted agents and, therefore, a better understanding of the underlying biology associated with these tumors beyond the mutation itself will be needed to gain knowledge on potential predictors of therapy response.

Overall, these studies support that the characterization of DNA alterations alone is not able to fully cope with the complexity and heterogeneity underlying NSCLC. In this context, a comprehensive stratification incorporating transcriptional profiling might help improving patient selection and delivery of the most suitable treatment or combination of treatments depending on each individual.

2.4 Integration of drug sensitivity data from large cancer cell lines pharmacogenomic studies for specific treatment strategies identification

In **section 2.3**, the fact that a non-negligible percentage of patients undergo incomplete responses or display no response at all to the prescribed treatments was highlighted. Reasons behind this variability may include different immune landscapes, tumor-microenvironment interactions, or deregulation of different transcriptional programs. In this context, the potential advantages of introducing whole transcriptome profiling into NSCLC clinical management to further understand the complexity of the disease and the potential reasons of drug response variability were discussed. However, to assess the impact of potential predictors on drug response it is crucial to study the associations between these molecular features and the sensitivity/resistance to specific therapeutic compounds. In this regard, the currently ongoing cancer cell line drug screening projects are paving the way towards more comprehensive approaches that involve the characterization of large collections of cancer cell lines in terms of their genomic changes; cellular states at the RNA, protein, and post-translational levels; and determining their sensitivities to anticancer drugs. With a sufficiently large collection of cell models, one can correlate therapeutic vulnerabilities with specific molecular characteristics, providing invaluable insights

into cancer biology, markers for patient selection, and potential new targets for cancer drug development (102,103).

In this chapter, the most prominent pharmacogenomics projects on cancer cell lines (CCLs) are reviewed, as well as the potential benefits of integrating these data into NSCLC precision medicine research.

2.4.1 Large cancer cell lines pharmacogenomic projects: CCLE, GDSC, CTRP, PRISM

Pan-cancer high-throughput drug screens are majorly performed on human CCLs, which are population of cells propagated in two-dimension *in vitro* culture (104). NCI-60 Human Tumors Cell Lines Screen (NCI60) was born in 1990 and has accumulated data on cell viability upon treatment with more than 50,000 drug compounds in 60 CCLs from nine different tissues (105). Although the number of drugs tested is huge, the reduced sample size may affect the robustness of the outcomes coming out from this study. In this context, the Cancer Cell Line Encyclopedia (CCLE), the Cancer Therapeutic Response Portal (CTRP), the Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) repurposing resource and the Genomics of Drug Sensitivity in Cancer (GDSC) were created to comprise a larger number of cell lines covering a wide range of tumour types (102,106–108). Currently, CCLE provides molecular data of almost 2,000 cancer cell lines, which include gene expression profiles, DNA alterations, copy number variants, DNA methylation profiles,

gene fusions, proteomic profiles, as well as genetic dependencies through short hairpin RNA (shRNA) and CRISPR-Cas9 knockdown screens. Moreover, CCLE and GDSC cell lines have been used by drug screening projects to generate drug sensitivity metrics for hundreds of cell lines and compounds. All these data have been compiled into the Dependency Map portal (DepMap, <https://depmap.org/portal/>), and can be easily downloaded for analysis (**Figure 13**) (109). Also, the DepMap portal provides with some interactive analysis tools that allow researchers to easily find associations between molecular features and gene dependencies or drug sensitivity for a specific user-defined list of cancer cell lines.

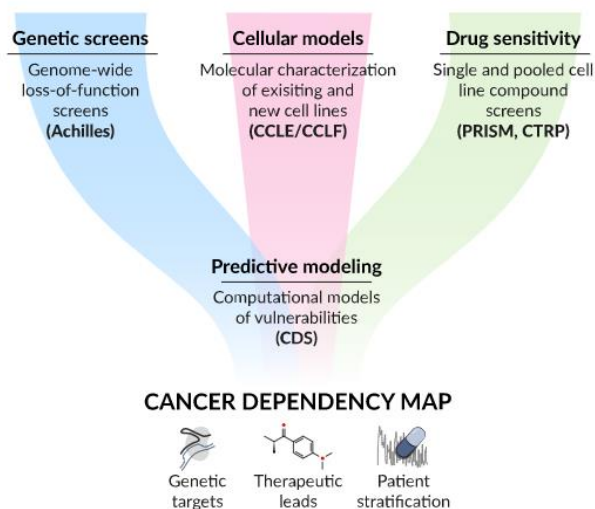


Figure 13. DepMap project data and aims.

Schema depicting the principal data sources included in the DepMap portal and the main objective of the project: combine different data layers to construct predictive models for new targets identification, propose new drugs and patient selection based on molecular profiles (Source: Extracted from (109))

2.4.2 Large cancer cell lines pharmacogenomic projects as a tool for bringing precision medicine to patients with cancer

Although nowadays there are numerous preclinical models available for cancer research (i.e., cancer cell lines, genetically engineered mouse models, patient derived xenografts (PDX) and three-dimensional culture systems or organoids), immortal cancer cell lines (CCLs) continue to be the most widely used model for the discovery of potential therapeutic vulnerabilities. Moreover, the generation of large pharmacogenomic studies, such as those described in section 4.1, has considerably accelerated the discovery of clinically relevant molecular features-drug associations and subsequent drug development (110). In this context, lung cancer is the most represented tumor type within DepMap repository accounting for around 23% of all tumor tissues within the dependency screens dataset (**Figure 14**). Moreover, most lung CCLs can develop tumors into animals for *in vivo* preclinical drug testing and validation (111).

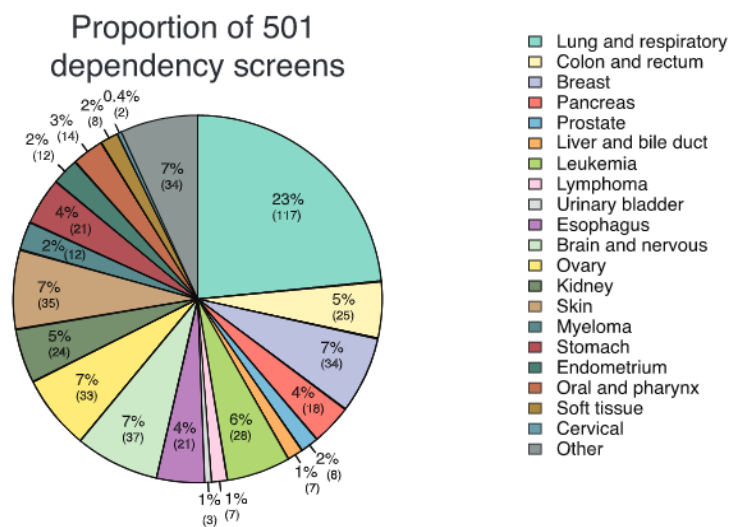


Figure 14. Tumor types by their cancer cell line representation in the dependency screens dataset.

(Source: Extracted from (111)).

One of the most common uses of CCLs has been to predict the efficacy of different therapeutic strategies. Models of LUAD have been widely used to assess sensitivity and resistance to different targeted therapies. For instance, several CCLs models with oncogenic *EGFR* mutations have been used to test different targeted therapies, derive gene expression signatures to predict the potential sensitivity or resistance to these strategies in the clinic and propose potential molecular mechanisms of resistance (69). Large scale pharmacogenomic projects could also be useful to improve patient stratification in the context of driver-negative NSCLC, which is a common scenario for LUSC patients. In this way, taking advantage of the genomic information for NSCLC CCLs

in the DepMap repository, one could first select driver-negative models and then integrate drug sensitivity and other molecular data, such as gene expression, to find potential therapeutic vulnerabilities and associated gene expression patterns and provide a biological rationale for guiding the antitumor treatment. Finally, NSCLC-CCLs molecular data could also be used to classify cell lines based on biological patterns previously derived from primary tumor profiles (i.e., gene expression, methylation) to identify potential treatments for those specific molecular subtypes.

Overall, these datasets constitute very important resources that allow researchers to link cancer-related molecular features with specific therapeutic vulnerabilities. The exploitation of these data may help improve NSCLC patient selection and, in the end, deliver effective treatments for specific NSCLC sub-populations.

3. HYPOTHESIS

Given the limitations of DNA alterations not only to capture the molecular complexity of tumors, but also to predict response to specific therapeutic strategies, innovative approaches are needed to improve patients' clinical outcomes. In this way, transcriptional profiling has already been used to further stratify colorectal cancer, breast cancer and NSCLC. More importantly, these subtypes were found to be associated with specific clinicopathological and molecular features. However, transcriptional subtypes have been established based on differential gene expression individual measures in NSCLC, which introduces variability between studies, leading to low overlap and lack of reproducibility. Moreover, the lack of associations between these subtypes and response to specific therapeutic treatments in most of the studies has hindered their clinical applicability. In this context, several large projects on pharmacogenomics performed in cancer cell lines (i.e., CCLE, GDSC, CTRP, and PRISM) has provided great amounts of drug sensitivity data to establish potential associations between tumor molecular alterations and specific treatments, which would be especially relevant for patients not harboring actionable genomic alterations.

The working hypothesis of this thesis is that the analysis of the pathway transcriptional footprint of LUAD and LUSC tumors will be able to stratify patients into molecularly different subtypes with potential implications on prognosis and the

response to specific treatment strategies. Overall, the proposed classification framework may delineate innovative therapeutic strategies beyond currently available DNA alteration-based targeted therapies, which is especially relevant in the case of driver-negative LUAD and LUSC patients, in which options beyond chemoimmunotherapy are very limited.

4. OBJECTIVES

4.1 General objectives

- **General objective 1:** Definition of transcriptional-based consensus molecular subtypes in lung adenocarcinoma and lung squamous cell carcinoma.
- **General objective 2:** Identification of potential vulnerabilities and drug candidates for the different lung adenocarcinoma or lung squamous cell carcinoma molecular subtypes.

4.2 Specific objectives

The two general objectives presented in 4.1, and which correspond to two separate publications, have common specific objectives which are listed hereafter:

- **Specific objective 1:** Identification, collection, and annotation of gene expression datasets of lung adenocarcinoma and lung squamous cell carcinoma human tumors.
- **Specific objective 2:** Quantification of the activity levels of a pre-defined list of fifty landmark molecular pathways across lung adenocarcinoma and lung squamous cell carcinoma tumor samples separately.
- **Specific objective 3:** Definition of consensus transcriptional subtypes based on the joint behavior of the evaluated signaling pathways in lung adenocarcinoma and lung squamous cell carcinoma.

- **Specific objective 4:** Clinical and molecular characterization of the identified lung adenocarcinoma and lung squamous cell carcinoma subpopulations (i.e., available clinical covariates, DNA alterations, mutational signatures, genome instability, immune landscape).
- **Specific objective 5:** Identification of specific therapeutic vulnerabilities for the subtypes using publicly available large-scale pharmacogenomic data from lung adenocarcinoma and lung squamous cell carcinoma cancer cell lines.

5. RESULTS

5.1 Publications

The two publications presented in this thesis have been published in international peer-reviewed journals.

The references are the following:

Hijazo-Pechero S, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Nadal E, Solé X. Transcriptional profiling of molecular pathways allows for the definition of robust lung squamous cell carcinoma molecular subtypes with specific vulnerabilities. *Clin Transl Med.* 2023 Sep;13(9):e1413. doi: 10.1002/ctm2.1413. PMID: 37735777; PMCID: PMC10514261.

Hijazo-Pechero S, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Montalban-Casafont A, Nadal E, Solé X. Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications. *Mol Oncol.* 2024 Feb;18(2):453-470. doi: 10.1002/1878-0261.13550. Epub 2023 Dec 21. PMID: 37943164; PMCID: PMC10850798.

5.2 Global results summary

5.2.1 Directors' report



Escola de Doctorat

MODEL Informe director/s /tutor sobre l'autorització del dipòsit de la tesi

Dr./a. Xavier Solé Acha , com a director de la tesi doctoral titulada “ Re-definition of non-small cell lung cancer transcriptional subtypes using integrative bioinformatics approaches” i, d'acord amb el que s'estableix a l'article 35 Normativa reguladora del Doctorat a la Universitat de Barcelona, emeto el següent:

INFORME

(Informe detallat i motivat sobre el contingut de la tesi i sobre l'autorització de dipòsit de la tesi que s'ha demanat)

El treball de tesi realitzat per la doctoranda Sara-Hijazo Pechero per optar al títol de doctora a la Universitat de Barcelona, s'ha dut a terme sota la meua co-direcció juntament amb la del Dr. Ernest Nadal a l'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL). La tesi es presenta com a compendi de dos articles:

1. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Nadal E, Solé X. Transcriptional profiling of molecular pathways allows for the definition of robust lung squamous cell carcinoma molecular subtypes with specific vulnerabilities. Clin Transl Med. 2023 Sep;13(9):e1413. doi: 10.1002/ctm2.1413. PMID: 37735777; PMCID: PMC10514261. **IF 2022: 10.6 (DECIL 1, MEDICINE, RESEARCH & EXPERIMENTAL)**
2. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Montalban-Casafont A, Nadal E, Solé X. Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications. Mol Oncol. 2024 Feb;18(2):453-470. doi: 10.1002/1878-0261.13550. Epub 2023 Dec 21. PMID: 37943164; PMCID: PMC10850798. **IF 2022: 6.6 (QUARTIL 1, ONCOLOGY)**

Atès que compleix tots els estàndards, puc afirmar que la memòria que es presenta té la qualitat suficient per ser defensada davant del tribunal corresponent.

Barcelona, 25 d'/de Març de 2024.

(signat)

Dr./a **Xavier Solé Acha**

Un cop s'hagi emplenat l'informe, s'ha d'adjuntar i s'ha de fer arribar al doctorand o al president de la Comissió Acadèmica del programa de doctorat responsable de la tesi.



MODEL Informe director/s /tutor sobre l'autorització del dipòsit de la tesi

Dr./a. Ernest Nadal , com a director de la tesi doctoral titulada “ Re-definition of non-small cell lung cancer transcriptional subtypes using integrative bioinformatics approaches” i, d'acord amb el que s'estableix a l'article 35 Normativa reguladora del Doctorat a la Universitat de Barcelona, emeto el següent:

INFORME

(Informe detallat i motivat sobre el contingut de la tesi i sobre l'autorització de dipòsit de la tesi que s'ha demanat)

La memòria que es presenta té la qualitat suficient per la seva defensa com a treball de Tesi doctoral. La tesi es presenta com a compendi de dos articles:

1. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Nadal E, Solé X. Transcriptional profiling of molecular pathways allows for the definition of robust lung squamous cell carcinoma molecular subtypes with specific vulnerabilities. Clin Transl Med. 2023 Sep;13(9):e1413. doi: 10.1002/ctm2.1413. PMID: 37735777; PMCID: PMC10514261. **IF: 10,6**
2. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Montalban-Casafont A, Nadal E, Solé X. Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications. Mol Oncol. 2024 Feb;18(2):453-470. doi: 10.1002/1878-0261.13550. Epub 2023 Dec 21. PMID: 37943164; PMCID: PMC10850798. **IF: 6,6**

En resum, els treballs realitzats per la doctoranda, Sara Hijazo Pechero, presenten un nou marc de classificació pels tumors de pulmó de cèl·lula no petita basat en el perfil d'expressió de 50 vies de senyalització. Aquest mètode resulta més sòlid que la classificació basada en els nivells d'expressió de gens individuals, propens a múltiples fonts de variabilitat. Els resultats demostren que els tumors de pulmó es classifiquen en diferents subtipus amb empremtes transcripcionals específiques. Aquests subgrups podrien tenir implicacions pel que fa a la resposta a diferents tractaments. A més, en aquest estudi s'observa que els grups s'associen amb determinades alteracions genòmiques i diferents patrons a nivell del microambient immune. D'altra banda, el treball proposa una prioritització de fàrmacs per cadascun dels subgrups, basant-se en les dades públiques farmaco-genòmiques generades en línies cel·lulars de càncer de pulmó de cèl·lula no petita. En general, aquest estudi podria millorar l'estratificació del pacients més enllà de la genòmica i dels biomarcadors individuals, posant el focus en l'anàlisi transcriptòmica. Aquesta aproximació podria revelar noves opcions de tractament més personalitzades en càncer de pulmó de cèl·lula no petita, sobretot en aquells pacients que no tenen cap alteració tractable.

Durant tot aquest temps, la doctoranda ha demostrat una gran capacitat de treball, ha desenvolupat capacitats suficients per desenvolupar les tasques de recerca assignades, així com el pensament crític, la interpretació i presentació dels resultats. A més durant el seu doctorat, ha col·laborat en diversos projectes de recerca amb altres investigadors. Tot plegat, puc afirmar que després de l'experiència del doctorat, la Sara Hijazo Pechero està preparada per continuar la seva carrera com investigadora i emprendre una nova etapa amb major autonomia.

Barcelona, 25 d'/de Març de 2024.



(signat)

Dr. Ernest Nadal Alforja

Un cop s'hagi emplenat l'informe, s'ha d'adjuntar i s'ha de fer arribar al doctorand o al president de la Comissió Acadèmica del programa de doctorat responsable de la tesi.

Escola de Doctorat

MODEL Informe director/s /tutor sobre l'autorització del dipòsit de la tesi

Dr./a. Víctor Moreno , com a tutor de la tesi doctoral titulada “ Re-definition of non-small cell lung cancer transcriptional subtypes using integrative bioinformatics approaches” i, d'acord amb el que s'estableix a l'article 35 Normativa reguladora del Doctorat a la Universitat de Barcelona, emeto el següent:

INFORME

(Informe detallat i motivat sobre el contingut de la tesi i sobre l'autorització de dipòsit de la tesi que s'ha demanat)

El treball de tesi realitzat per la doctoranda Sara-Hijazo Pechero per optar al títol de doctora a la Universitat de Barcelona, té la qualitat suficient per ser defensat davant del tribunal corresponent. La tesi es presenta com a compendi de dos articles:

1. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Nadal E, Solé X. Transcriptional profiling of molecular pathways allows for the definition of robust lung squamous cell carcinoma molecular subtypes with specific vulnerabilities. Clin Transl Med. 2023 Sep;13(9):e1413. doi: 10.1002/ctm2.1413. PMID: 37735777; PMCID: PMC10514261. **IF: 10.6**
2. **Hijazo-Pechero S**, Alay A, Cordero D, Marín R, Vilariño N, Palmero R, Brenes J, Montalban-Casafont A, Nadal E, Solé X. Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications. Mol Oncol. 2024 Feb;18(2):453-470. doi: 10.1002/1878-0261.13550. Epub 2023 Dec 21. PMID: 37943164; PMCID: PMC10850798. **IF: 6,6**

Com a tutor, autoritzo el dipòsit de la tesi, reconeixent el treball excepcional del doctorand.

Barcelona, 25 d'/de Març de 2024.



(signat)

Dr./a Víctor Moreno

Un cop s'hagi emplenat l'informe, s'ha d'adjuntar i s'ha de fer arribar al doctorand o al president de la Comissió Acadèmica del programa de doctorat responsable de la tesi.

5.2.2 Article 1: Transcriptional profiling of molecular pathways allows for the definition of lung squamous cell carcinoma molecular subtypes with specific vulnerabilities

In this work, we developed a gene expression-based classification of LUSC to improve patient stratification beyond histological and genomic features. For this purpose, we integrated gene expression profiles from more than 2,000 publicly available LUSC tumors. To our knowledge, this number of cases far exceeds the sample size used for all previously proposed LUSC transcriptional-based classifications. In addition, instead of relying on gene expression measures of individual genes, we evaluated the activity of 50 landmark molecular pathways. The evaluation of lists of genes that are coordinately expressed to constitute pathways is less prone to stochastic variations than single gene expression values. As a result, LUSC samples were classified in five transcriptional-based subtypes depending on the joint behavior of the studied pathways. These subtypes were characterized at different levels (i.e., clinical covariates, genomic features, immune landscape) and were validated using an independent dataset of LUSC tumors. Finally, the integration of publicly available cancer cell lines pharmacogenomic data suggested specific pharmacologic interventions for the subtypes, which are in line with their signaling pathway footprint.

The main ideas from our manuscript are:

- Current LUSC clinical management is unable to cope with disease complexity and predict drug response.
- Transcriptional profiling of landmark molecular pathways in a large LUSC cohort allows for the definition of five subtypes: SCC1 (9.9% of patients), SCC2 (23.9%), SCC3 (25.8%), SCC4 (31.0%) and SCC5 (9.4%).
- SCC1 and SCC4 subtypes correlated with higher genome instability, cell cycle-related pathway activity levels, and specific sensitivity to chemotherapy, based on LUSC cell lines data.
- These transcriptional subtypes have differential immune landscapes. SCC2 and SCC3 showed higher infiltration of a wide variety of immune cell populations and markers (i.e., PD-L1).
- Our SCC group definition was compared to previous classifications and validated using an independent dataset that was not included in the initial analysis.
- These results warrant further validation and might be useful for patients with lung LUSC, who have lower treatment options due to the lack of actionable driver alterations.

Transcriptional profiling of molecular pathways allows for the definition of robust lung squamous cell carcinoma molecular subtypes with specific vulnerabilities

Dear Editor,

Lung squamous cell carcinoma (SCC) is a histological subtype of non-small cell lung cancer associated with poor prognosis. Actionable driver alterations are extremely rare in SCC and standard of care (SoC) consists of immunotherapy alone or combined with chemotherapy based on the PD-L1 expression, with few long-term survivors. We hypothesized that transcriptomic data analysis can improve patient stratification and may unravel novel treatment approaches for these patients.¹

We developed a bioinformatics pathway-based classification framework using publicly available whole-transcriptome data from more than 2,000 SCC samples focusing on 50 pathways. Previous transcriptome-based classifications used individual gene expression measures, which are prone to multiple sources of variability.² Detailed methodology, gene expression datasets and schematic view of the bioinformatics framework are shown in Table S1 and Figures S1 and S2, respectively. Five SCC subtypes were identified based on the combined transcriptional behaviour of the 50 pathways (Figure 1A,B): SCC1 (9.9% of patients), SCC2 (23.9%), SCC3 (25.8%), SCC4 (31.0%) and SCC5 (9.4%). SCC subtypes displayed their specific transcriptional footprint, which could shape different treatment responses (Figure 1C and Figure S3). SCC1 and SCC4 showed higher activation levels of cell proliferation and DNA damage response (DDR) pathways, and rather low transcriptional activation levels of immune-related programs, especially SCC4. In contrast, SCC2 showed higher immune system-related pathways activation and lower cell cycle signatures activation. SCC3 exhibited high activity of proliferation and immune-related pathways, and upregulation of KRAS, NFKB, IL2-STAT5 and TNFA signaling pathways, which may play a role in shaping these tumours' immunity. SCC5 displayed reduced activation of most pathways. Our classification partly overlaps with previous intrinsic subtypes described by Wilkerson et al,³

where primitive subtype correlates with the proliferative SCC4, while secretory subtype was distributed between the immune-enriched SCC2 and SCC3, and the classical subtype overlapped with SCC3 and SCC4 (Figure S4). No significant differences were observed in clinicopathological characteristics or overall survival (Table S2 and Figure S5). This underlines the importance of this work aiming to distinguish consistent groups within an apparently clinically homogeneous population of patients with SCC and unravel potential transcriptional vulnerabilities.

SCC subtypes were further characterized to identify differential genomic patterns (Figure 2). For mutational signatures, tobacco-related genomic signature was found to be overrepresented across all subtypes, suggesting an equivalent tobacco-related DNA damage (Figure S6). Although subtle differences were observed regarding tumour mutational burden (TMB), we found higher copy number alterations (CNA) in SCC1 and SCC4 (Figure 2A,B and Tables S3 and S4). SCC1 further demonstrated higher DDR deficiency scores compared to other subtypes (Figure 2C and Table S5). Thus, the greater genomic instability found for SCC1 and SCC4 might not be the result of higher exposure to exogenous carcinogens (i.e. tobacco), but rather a consequence of DDR mechanisms, or replication stress.⁴ This classification framework and the reported association of the subtypes with genomic instability (i.e. higher CNA rates) was validated in an independent dataset of SCC (Figure 3).⁵ All five subtypes were found in the CPTAC-3 dataset and samples map within one of the consensus subtypes, which supports the robustness and reproducibility of this classification (Figure 3A). We observed that both CNA and pathways activation were concordant in both discovery and validation sets, therefore genomic alterations were consistent beyond the expression patterns used to classify these samples (Figure 3B,C).

Immune checkpoint inhibitors (ICI) alone, combined with chemotherapy, or following chemoradiotherapy are

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Clinical and Translational Medicine* published by John Wiley & Sons Australia, Ltd on behalf of Shanghai Institute of Clinical Bioinformatics.

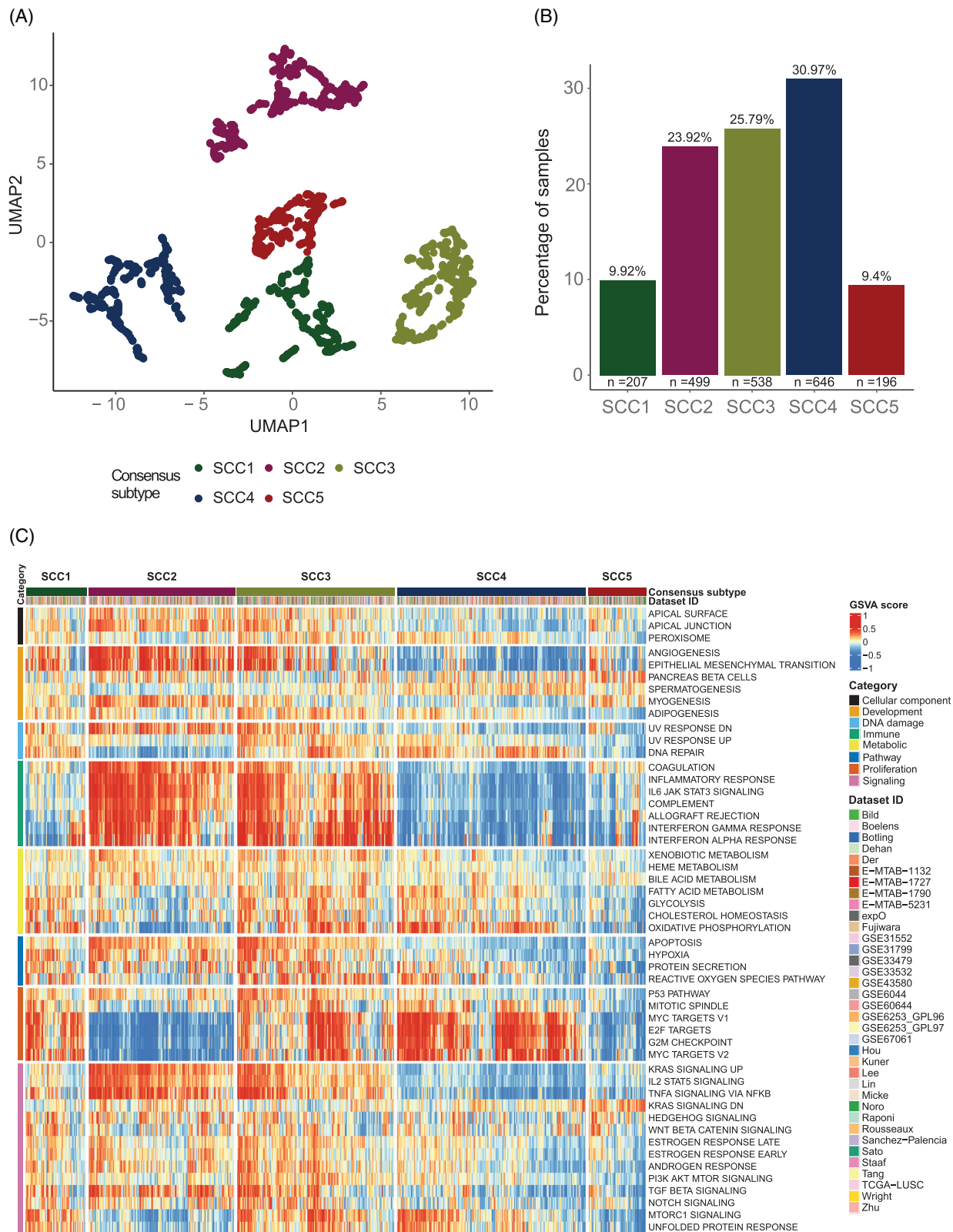


FIGURE 1 Overview of squamous cell carcinoma (SCC) groups at the transcriptional level. (A) Final consensus map of lung SCC tumours. Each dot represents the summary centroid of the different subpopulations identified during the classification process. Using UMAP and walktrap clustering method with Euclidean distance on these centroids, five different consensus subtypes represented by different colours were identified based on the joint behaviour of the 50 studied molecular pathways. (B) Distribution of the five identified lung SCC subtypes. (C) Relative activity levels of the 50 studied pathways in each of the 2086 SCC samples were assigned to a consensus subtype. Red colours indicate higher relative activity of a pathway in a certain sample, whereas blue colours indicate lower relative activity of a pathway in a certain sample.

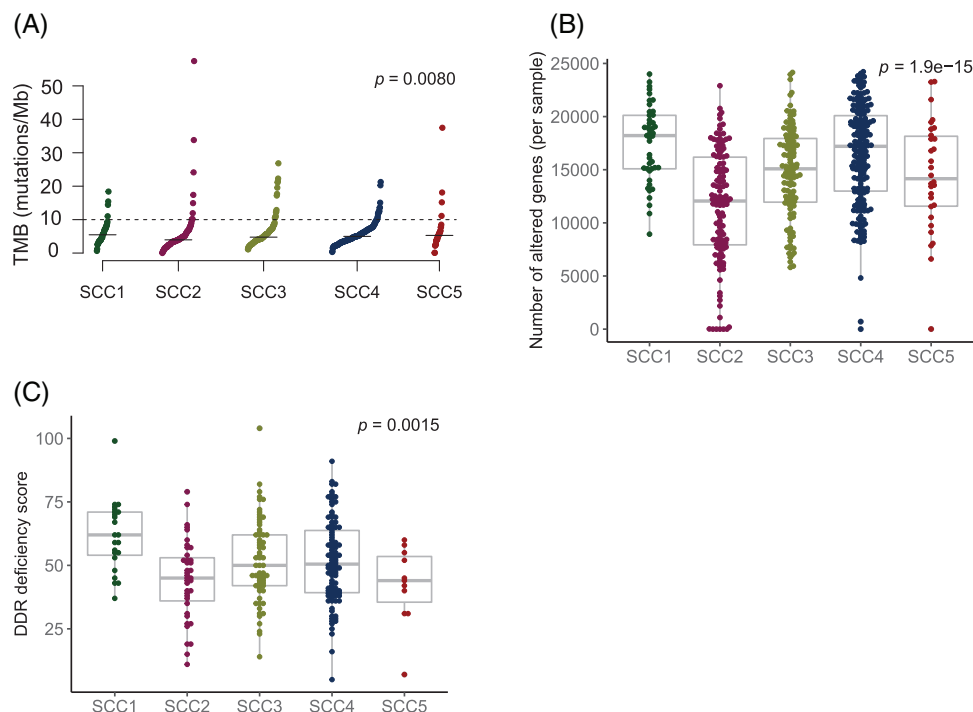


FIGURE 2 Genomic characterization. (A) Tumor mutational burden (TMB) across lung squamous cell carcinoma (SCC) consensus subtypes. Each dot represents the TMB value for a specific sample. The black segment represents the median TMB value for each lung SCC subtype. Kruskal-Wallis tests were used to make comparisons between groups. p-Value was corrected using the false discovery rate (FDR) multiple-comparisons correction method. (B) Copy number burden across lung SCC subtypes. Each dot represents the number of altered genes per sample. Kruskal-Wallis tests were used to make comparisons between groups. p-Value was corrected using the false discovery rate (FDR) multiple-comparisons correction method. (C) DNA damage repair (DDR) deficiency score distribution across lung SCC subgroups. Each dot represents the DDR score per sample. Kruskal-Wallis tests were used to make comparisons between groups. p-Value was corrected using the false discovery rate (FDR) multiple-comparisons correction method.

part of the SoC for advanced SCC.⁶ However, patient selection strategies, based on single biomarkers (i.e. TMB and PD-L1), fail to predict long-term clinical benefit in SCC.⁷ We evaluated immune-cell-specific signatures and immune-related gene expression, which revealed different immune landscapes for the subtypes, with potential clinical implications (Figure 4). For instance, SCC2 and SCC3 demonstrated higher infiltration for both anti-tumour (i.e., cytotoxic, CD8+ and T-helper 1 cells), and immunosuppressive populations (i.e. M2-macrophages and T-regulatory cells), which could eventually prevent an effective immune response (Figure 4A and Figure S7). SCC2 and SCC3 also comprised tumours with high expression of most ICI, including *CD274* (Figure 4B and Figure S8). Although further validation (i.e. scRNA-Seq) would be needed, these results highlight the need to characterize the immune contexture, along with conventional single biomarkers, to stratify patients and deliver tailored and effective treatment strategies for advanced SCC.

Chemotherapy is also a key treatment for patients with SCC. Understanding chemotherapy response patterns and improving patients' selection remains crucial. Integrative analysis of pharmacogenomic data in SCC cell lines (SCC-

CCL), showed that the SCC4 subtype might benefit from different chemotherapy regimens (i.e. average AAC above 0.5 in at least two studies), which correlates with the proliferative nature and higher genome instability observed for this subtype (Figure 4C). In this set of SCC cell lines, platinum-based agents showed AAC values below < 0.2, regardless of the SCC subtype, suggesting lower sensitivity to these compounds (data not shown). However, the evaluation of gene-expression signatures predicting platinum resistance showed that primary tumors classified as SCC4 and SCC5 would potentially be more sensitive to these chemotherapies (Figure S9).⁸ Moreover, SCC4 and SCC1 SCC-CCL showed potential sensitivity for some cell cycle and DNA damage-targeted therapies (Figure 4C).

Single biomarkers and individual gene signatures have shown a limited ability to capture tumour heterogeneity. No pathway was exclusively expressed in one of the subtypes, but their combined expression pattern allowed the identification of subtypes with unique transcriptional footprints. To simplify the classification framework, we derived gene expression signatures for each subtype in the discovery cohort and validated them in the validation cohort (Figure S10 and Table S6). However, these

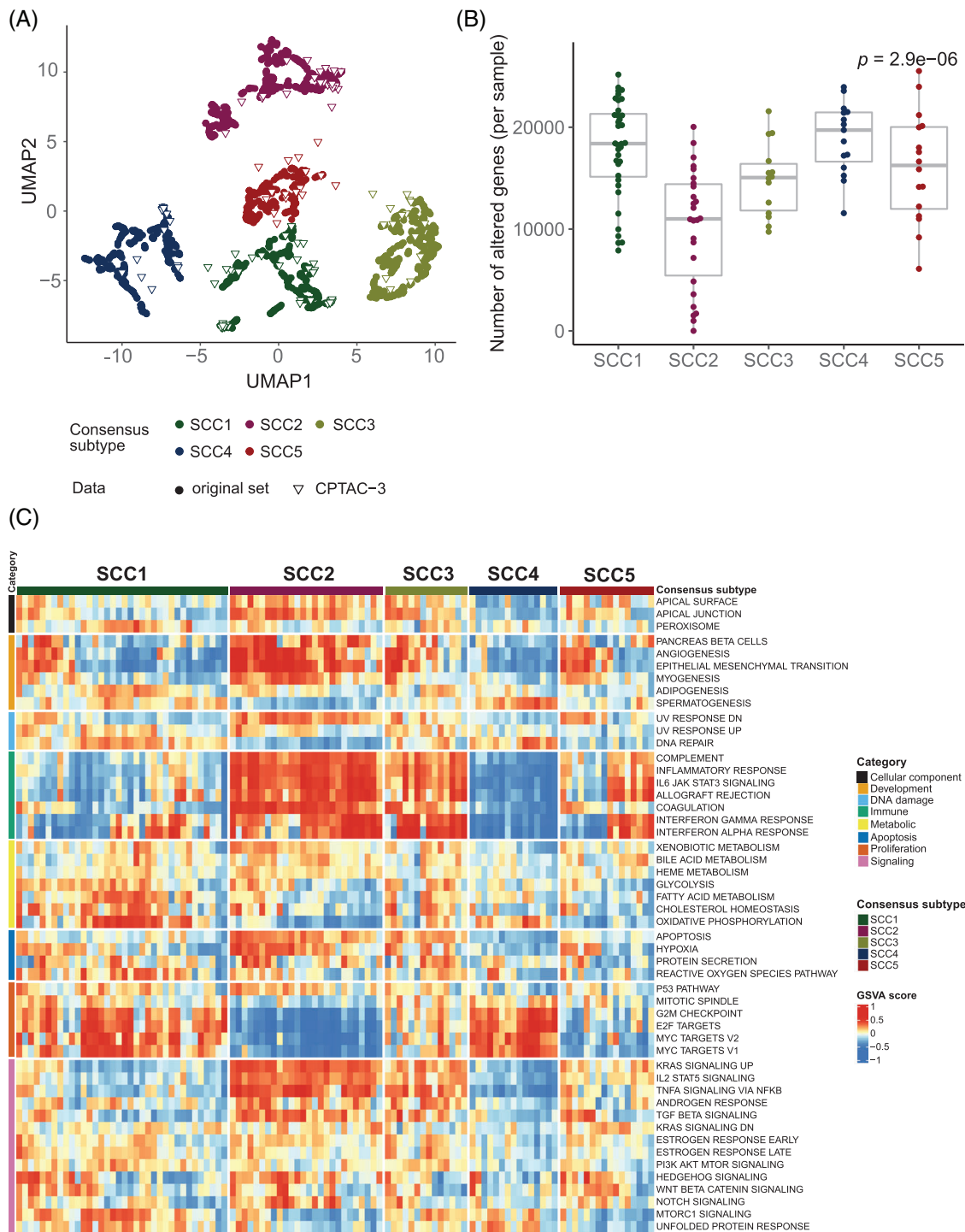


FIGURE 3 Lung squamous cell carcinoma (SCC) consensus subtypes independent validation. (A) New CPTAC-3 lung SCC samples were mapped on the previously established classification of SCC tumours based on the activity levels of the same 50 pathways used to define the original SCC subtypes. The new sample subtype status was decided based on the most frequent label of the closest neighbours of the original classification. Coloured circles represent samples used in the original set, whereas triangles represent new CPTAC-3 validation set samples. (B) Copy number burden across newly classified CPTAC-3 lung SCC samples. Each dot represents the number of altered genes per sample. The Kruskal-Wallis test was used to make comparisons between groups. *p*-Value was corrected using the false discovery rate (FDR) multiple-comparisons correction method. (C) Relative activity levels of the 50 studied pathways in each of the 108 CPTAC-3 lung SCC samples were assigned to a consensus subtype. Red colours indicate higher relative activity of a pathway in a certain sample, whereas blue colours indicate lower relative activity of a pathway in a certain sample.

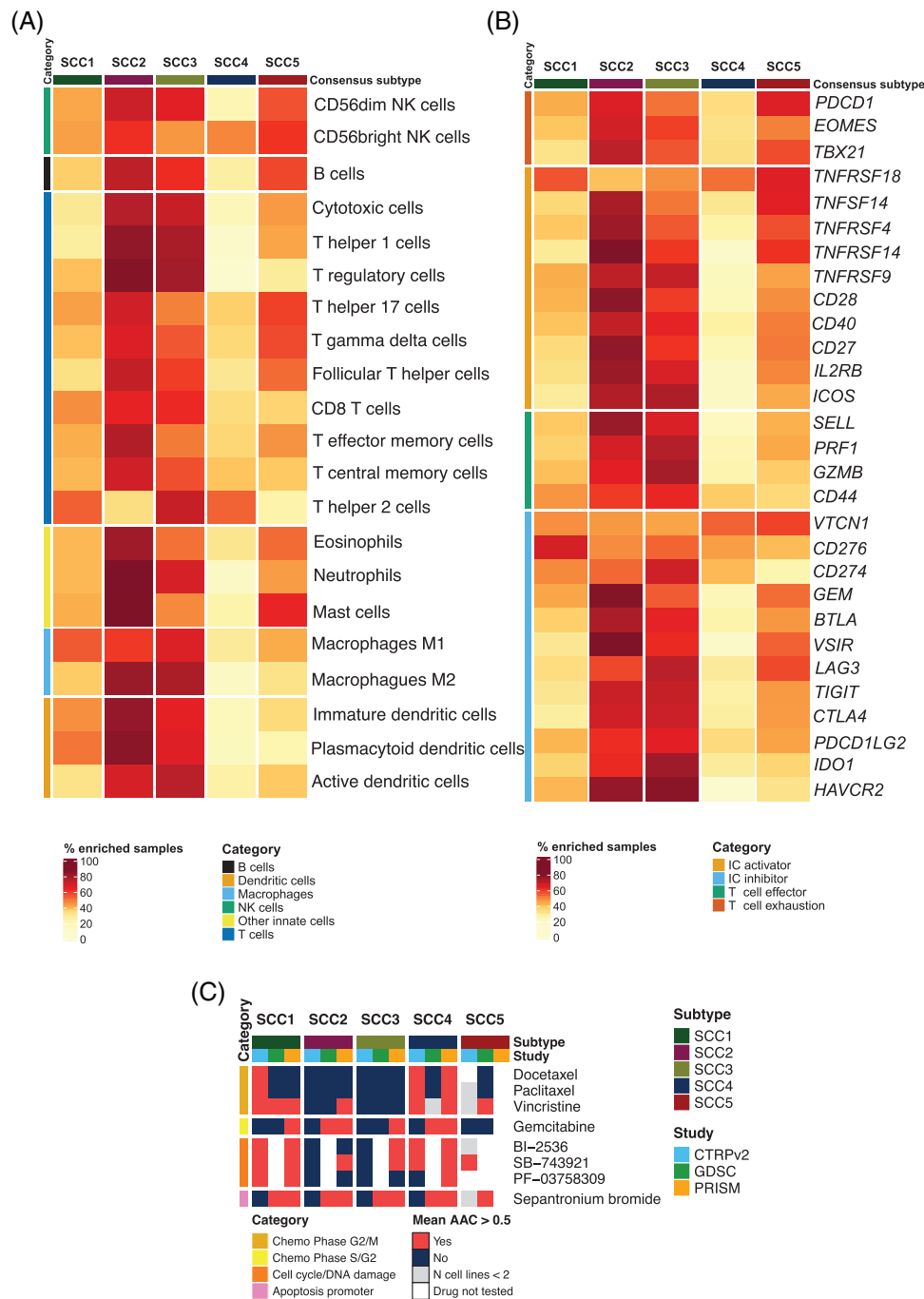


FIGURE 4 Immune contexture characterization and potential subtype-specific vulnerabilities. (A) Percentage of samples with high infiltration of each of the 21 evaluated immune cell types. Median immune cell abundance GSVA score values were used as a cut-off to designate if a sample is enriched for a specific immune cell. Different immune cell categories are represented with different colours on the left side of the heatmap. (B) Percentage of samples with high expression of each of the evaluated immune-related biomarkers. Median gene expression values for each gene in each gene expression dataset were used as a cut-off to designate if a sample is enriched for a specific biomarker. Different immune marker categories are represented with different colours on the left side of the heatmap. (C) Potential therapeutic vulnerabilities for the lung squamous cell carcinoma (SCC) subtypes based on CTRPv2, GDSC and PRISM lung SCC cell lines drug sensitivity data. Heatmap representing drugs with mean AAC values greater than 0.5 in at least two studies within the same subtype. Mean AAC values were only calculated if the drug had been tested in at least two different SCC-CCLs within a subtype and study. Subtypes were considered potentially sensitive to the treatment if the average AAC value for the cell lines classified within a certain group was greater than 0.5 for at least two out of the three evaluated pharmacogenomics studies.

signatures are still complex, and their discrimination ability needs to be further evaluated. We would rather embrace a whole transcriptome technology, applicable at the clinical level (i.e. HTG Edge-Seq), that enables the evaluation of the combined activity levels of the proposed pathways.

In conclusion, we have presented a comprehensive molecular classification of SCC, based on the transcriptional activity of 50 pathways. Although further validation is required, these results could be useful for improving precision medicine for patients with lung SCC, who have limited treatment options and heterogeneous responses to standard treatments.

ACKNOWLEDGEMENTS

We sincerely thank David G. Beer, Cristina Muñoz-Pinedo and the reviewers and editors for useful comments that significantly helped to improve the present manuscript. Copyediting editorial support was provided by Aurora O'Brate.

CONFLICT OF INTEREST STATEMENT

Xavier Solé participated in lectures from Roche. Ernest Nadal received research support from Roche, Pfizer, Merck-Serono and Bristol Myers Squibb; and participated in advisory boards or lectures from Bristol Myers Squibb, Merck Serono, Merck Sharpe & Dohme, Lilly, Roche, Pfizer, Bayer, Sanofi, Takeda, Boehringer Ingelheim, Janssen, Daiichi Sankyo, Amgen and AstraZeneca. The rest of the authors declare no conflict of interest.

FUNDING INFORMATION

Sara Hijazo-Pechero is supported by an AGAUR-FI fellowship (2022 FI_B2 00066), with the support of the FI program of the Secretariat for Universities and Research of the Department of Business and Knowledge of the Government of Catalonia, and the support of the European Union through the European Social Fund "ESF, Investing in your future". Xavier Solé received support from Ministerio de Ciencia, Innovación y Universidades, which is part of Agencia Estatal de Investigación (AEI), through Retos Research Grant, number RTI2018-102134-A-I00. (Co-funded by the European Regional Development Fund. ERDF, a way to build Europe). Ernest Nadal received support from Instituto de Salud Carlos III (grants PI18/00920 and PI21/00789) (co-funded by the European Regional Development Fund. ERDF, a way to build Europe). We thank the CERCA Programme / Generalitat de Catalunya for institutional support. Raúl Marín is supported with the funding of the Ministerio de Universidades, through the predoctoral fellowship number FPU19/01734 for the Formación de Profesorado Universitario (FPU). Noelia Vilariño and Jesús Brenes are supported by a Rio Hortega

contract (CM19/00245 & CM21/00073, respectively) from the Instituto de Salud Carlos III.

DATA AVAILABILITY STATEMENT

Gene expression data for SCC consensus classification were obtained from GEO and ArrayExpress public repositories. Specific information and identifiers of each dataset are available in Table S1.

CCLC and GDSC lung SCC cancer cell lines molecular data was obtained from <https://depmap.org> and https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html, respectively. Specific cell line identifiers and sources are detailed in the Supporting Information.

GDSC, CTRPv2 and PRISM studies drug sensitivity data were available within the *PharmacoGx* R package.

Lung SCC CPTAC-3 study gene expression and copy number alterations data were downloaded from the Supporting Information.⁵

Sara Hijazo-Pechero^{1,2,3}

Ania Alay^{1,2}

David Cordero^{1,2}

Raúl Marín^{1,2}

Noelia Vilariño^{2,4,5}

Ramón Palmero^{2,4}

Jesús Brenes⁴ 

Ernest Nadal^{2,4}

Xavier Solé^{3,6} 

¹Unit of Bioinformatics for Precision Oncology, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Spain

²Preclinical and Experimental Research in Thoracic Tumors (PrETT), Molecular Mechanisms and Experimental Therapy in Oncology Program (Oncobell), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Spain

³Translational Genomics and Targeted Therapies in Solid Tumors, August Pi i Sunyer Biomedical Research Institute (IDIBAPS), Barcelona, Spain

⁴Thoracic Oncology Unit, Catalan Institute of Oncology (ICO), Barcelona, Spain

⁵Neuro-Oncology Unit, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Spain

⁶Molecular Biology CORE, Center for Biomedical Diagnostics (CDB), Hospital Clínic de Barcelona, Barcelona, Spain

Correspondence

Xavier Solé, Molecular Biology CORE, Center for Biomedical Diagnostics (CDB), Hospital Clínic de

Barcelona, C. de Villarroel 170, stair 5 floor 5, 08036
Barcelona, Spain.
Email: xasole@clinic.cat

Ernest Nadal, Department of Medical Oncology, Catalan
Institute of Oncology, Avda Gran Via de L'Hospitalet
199–203, L'Hospitalet de Llobregat, Barcelona, Spain.
Email: esnadal@iconcologia.net

ORCID

Jesús Brenes  <https://orcid.org/0000-0003-0238-3319>

Xavier Solé  <https://orcid.org/0000-0002-2197-3325>

REFERENCES

1. Hendriks LE, Kerr KM, Menis J, et al. Non-oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol*. 2023;34(4):358-376. doi:[10.1016/j.annonc.2022.12.013](https://doi.org/10.1016/j.annonc.2022.12.013)
2. Hijazo-Pechero S, Alay A, Marín R, et al. Gene expression profiling as a potential tool for precision oncology in non-small cell lung cancer. *Cancers*. 2021;13(19):4734. doi:[10.3390/cancers13194734](https://doi.org/10.3390/cancers13194734)
3. Wilkerson MD, Yin X, Hoadley KA, et al. Lung squamous Cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res*. 2010;16(19):4864-4875. doi:[10.1158/1078-0432.CCR-10-0199](https://doi.org/10.1158/1078-0432.CCR-10-0199)
4. Zhu H, Swami U, Preet R, Zhang J. Harnessing DNA replication stress for novel cancer therapy. *Genes*. 2020;11(9):990. doi:[10.3390/genes11090990](https://doi.org/10.3390/genes11090990)
5. Satpathy S, Krug K, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell*. 2021;184(16):4348-4371. doi:[10.1016/j.cell.2021.07.016](https://doi.org/10.1016/j.cell.2021.07.016) e40
6. Ettinger DS, Aisner DL, Wood DE, et al. NCCN guidelines insights: non-small cell lung cancer, version 5.2018. *J Natl Compr Cancer Netw JNCCN*. 2018;16(7):807-821. doi:[10.6004/jnccn.2018.0062](https://doi.org/10.6004/jnccn.2018.0062)
7. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Updated analysis of KEYNOTE-024: Pembrolizumab versus platinum-based chemotherapy for advanced non-small-cell lung cancer with PD-L1 tumor proportion score of 50% or greater. *J Clin Oncol Off J Am Soc Clin Oncol*. 2019;37(7):537-546. doi:[10.1200/JCO.18.00149](https://doi.org/10.1200/JCO.18.00149)
8. Mucaki EJ, Zhao JZL, Lizotte DJ, Rogan PK. Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. *Signal Transduct Target Ther*. 2019;4(1):1-12. doi:[10.1038/s41392-018-0034-5](https://doi.org/10.1038/s41392-018-0034-5)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

5.2.3 Article 2: Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications



In this work we integrated the transcriptional profiles of more than 4,500 LUAD and, based on the activity levels of a set of 50 molecular pathways, we were able to identify seven LUAD molecular subtypes. Importantly, the number of samples included in this study further exceeds that of previous studies, covering an important part of the molecular diversity of LUAD. This classification was associated with survival outcomes and was correlated with relevant clinical characteristics. Besides, at the genomic level, LUAD transcriptional subtypes were associated with the presence of oncogenic driver alterations, mutational signatures, copy number alterations burden and DNA damage repair capacity. Furthermore, the integration of drug sensitivity data from three large pharmacogenomics studies unraveled potential therapeutic vulnerabilities for the subtypes. Finally, the transcriptional subtypes showed distinct patterns in terms of immune cells infiltration and immune-related biomarkers expression and were able to predict immune response in addition to PD-L1 gene expression and TMB markers.

Main results from this work are:

- Current LUAD clinical management is unable to cope with disease complexity and predict drug response.
- Transcriptional profiling of landmark molecular pathways in a large LUAD cohort allows for the definition of seven subtypes: AD1 (16.84% of patients), AD2 (18.70%), AD3 (10.43%), AD4 (13.19%), AD5 (20.95%), AD6 (17.06%) and AD7 (2.86%). This classification was successfully validated in an independent cohort.
- AD1, AD4 and AD5 subtypes were associated with better overall survival.
- AD2, AD3, AD6 and AD7 were associated with worse overall survival.
- AD1 and AD4 subtypes were enriched in EGFR mutations, whereas AD2 and AD6 showed higher TP53 alteration frequencies. Tumors with the same oncogenic alteration were classified in the same subtype, underscoring the importance of our classification to explain LUAD molecular heterogeneity beyond genomic classification. Conversely, the fact that tumors with different driver alterations coexist in the same transcriptional subtype, suggests that different oncogenic mutations may give rise to similar transcriptional phenotypes, which could benefit from similar combinatorial strategies.
- AD2 and AD6 subtypes correlated with higher genome instability, proliferation-related pathways expression and specific sensitivity to chemotherapy, based on LUAD cell lines data.

- Pathway transcriptional profiling-based LUAD subtypes were able to predict immunotherapy response in addition to PD-L1 expression and TMB biomarkers. Tumors within AD4 subtype were found to be 2.9 times more likely to respond to immunotherapy compared to the tumors classified in any other subtype. Despite being among the most infiltrated subtypes and showing high PD-L1 expression, only 12.5% of AD3 tumors were predicted as potential responders. Also, in correlation with its immune excluded phenotype, AD2 tumors were 80% less likely to respond to immunotherapy than other subtypes, based on *in silico* predictions.

Transcriptional analysis of landmark molecular pathways in lung adenocarcinoma results in a clinically relevant classification with potential therapeutic implications

Sara Hijazo-Pechero^{1,2,3}, Ania Alay^{1,2}, David Cordero^{1,2}, Raúl Marín^{1,2}, Noelia Vilariño^{2,4,5}, Ramón Palmero^{2,4}, Jesús Brenes⁴, Aina Montalban-Casafort⁶, Ernest Nadal^{2,4}  and Xavier Solé^{3,6} 

1 Unit of Bioinformatics for Precision Oncology, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain

2 Preclinical and Experimental Research in Thoracic Tumors (PrETT), Molecular Mechanisms and Experimental Therapy in Oncology Program (Oncobell), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

3 Translational Genomics and Targeted Therapies in Solid Tumors, August Pi i Sunyer Biomedical Research Institute (IDIBAPS), Barcelona, Spain

4 Thoracic Oncology Unit, Department of Medical Oncology, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain

5 Neuro-Oncology Unit, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain

6 Molecular Biology CORE, Center for Biomedical Diagnostics (CDB), Hospital Clínic de Barcelona, Spain

Keywords

immunotherapy; lung adenocarcinoma; molecular pathways; precision oncology; transcriptional subtypes

Correspondence

X. Solé, Molecular Biology CORE, Center for Biomedical Diagnostics (CDB), Hospital Clínic de Barcelona, C. de Villarroel 170, stair 5 floor 5, 08036 Barcelona, Spain
Tel: +34 932 275 400 ext. 1798
E-mail: xasole@clinic.cat

and

E. Nadal, Catalan Institute of Oncology (ICO), Preclinical and Experimental Research in Thoracic Tumors, Bellvitge Biomedical Research Institute (IDIBELL), Avinguda de la Gran Via de l'Hospitalet 199-203, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
Tel: +34 932 60 77 44
E-mail: esnadal@iconcologia.net

(Received 9 May 2023, revised 11 September 2023, accepted 3 November 2023, available online 21 December 2023)

doi:10.1002/1878-0261.13550

Lung adenocarcinoma (LUAD) is a molecularly heterogeneous disease. In addition to genomic alterations, cancer transcriptional profiling can be helpful to tailor cancer treatment and to estimate each patient's outcome. Transcriptional activity levels of 50 molecular pathways were inferred in 4573 LUAD patients using Gene Set Variation Analysis (GSVA) method. Seven LUAD subtypes were defined and independently validated based on the combined behavior of the studied pathways: AD (adenocarcinoma subtype) 1–7. AD1, AD4, and AD5 subtypes were associated with better overall survival. AD1 and AD4 subtypes were enriched in epidermal growth factor receptor (*EGFR*) mutations, whereas AD2 and AD6 showed higher tumor protein p53 (*TP53*) alteration frequencies. AD2 and AD6 subtypes correlated with higher genome instability, proliferation-related pathway expression, and specific sensitivity to chemotherapy, based on data from LUAD cell lines. LUAD subtypes were able to predict immunotherapy response in addition to *CD274* (PD-L1) gene expression and tumor mutational burden (TMB). AD2 and AD4 subtypes were associated with potential resistance and response to immunotherapy, respectively. Thus, analysis of transcriptomic data could improve patient stratification beyond genomics and single biomarkers (i.e., PD-L1 and TMB) and may lay the foundation for more personalized treatment avenues, especially in driver-negative LUAD.

Abbreviations

AAC, area above the curve; CNA, copy number alteration; DDR, DNA damage repair; FDR, false discovery rate; GEO, Gene Expression Omnibus; GSVA, gene set variation analysis; ICA, immune checkpoint activators; ICI, immune checkpoint inhibitors; LUAD, lung adenocarcinoma; LUAD-CCL, lung adenocarcinoma cancer cell lines; NSCLC, non-small cell lung cancer; OS, overall survival; SNV, somatic single nucleotide variants; TMB, tumor mutational burden; TPM, transcripts per million; UMAP, Uniform Manifold Approximation and Projection; WES, whole-exome sequencing; WHO, World Health Organization.

1. Introduction

Lung cancer is a major global health problem. According to the World Health Organization (WHO), lung cancer was the leading cause of cancer-related deaths and the second most frequently diagnosed cancer in 2020 [1]. Regarding histological subtypes, lung adenocarcinoma (LUAD) is the most prevalent histological entity, accounting for almost 55% of the diagnoses [2]. In terms of clinical management, chemotherapy alone or in combination with immunotherapy is considered the standard of care for patients with advanced LUAD not harboring actionable oncogenic alterations [3]. Additionally, recent advances in high-throughput genomic technologies for molecular profiling have accelerated the evolution of personalized medicine [4,5]. For instance, the current management of LUAD requires molecular testing to detect actionable genomic alterations predicting clinical benefit to targeted therapies [3]. However, patients with advanced LUAD have heterogeneous responses and poor survival outcomes (5-year survival rate = 21%) [6]. These differences between patient response rates have been attributed to tumor burden, comorbidities, functional status, or tumor heterogeneity, such as different immune landscapes, activation of signaling pathways, and presence of different cell types [7]. Thus, improving LUAD patients' stratification beyond genomic testing could move forward precision medicine, but is a major challenge.

Given the limitations of genomics to capture the complexity of LUAD and to predict response to specific treatments, innovative approaches are needed to improve clinical outcome. In this regard, gene expression profiling has already been used to further stratify LUAD into different molecular subtypes [8]. However, the clinical relevance of those classifications was questioned due to technical intrinsic limitations, inconsistencies between studies, and the lack of association with potential therapeutic strategies.

The aim of our study was to develop a novel LUAD classification based on transcriptomics able to improve patients' stratification beyond the current histological and genomic-based classifications. For this purpose, we integrated transcriptional profiles from more than 4500 LUAD. To the best of our knowledge, this is the largest study defining transcriptional LUAD subtypes [8]. In addition, unlike previous attempts relying on measuring individual gene expression, we assessed the activity of a set of well-defined molecular pathways, which makes it less prone to variability [9,10]. Based on this, a computational framework was developed to stratify LUAD into different subtypes based on the

expression of specific signaling pathways. These subtypes were further characterized at different levels (i.e., clinical covariates, genomic features, and immune landscape). Finally, the analysis of publicly available large-scale cancer cell line drug screening projects revealed potential therapeutic vulnerabilities for each group of LUAD tumors [11–14]. Overall, this classification may delineate novel therapeutic strategies beyond current genomic-based targeted therapies, which could be especially relevant in the case of driver-negative LUAD patients.

2. Materials and methods

2.1. Datasets and gene expression data processing

LUAD transcriptional profiles were obtained from Gene Expression Omnibus (GEO), Lung Cancer Explorer, and ArrayExpress public data archives [15–17]. Subsequent filters were applied to keep human LUAD tumor samples, exclude datasets with less than 10 samples, and remove those studies using platforms that do not cover a significant part of the transcriptome (i.e., targeted panels covering a smaller subset of genes). Overall, 56 datasets were included in this analysis, constituting more than 4500 LUAD samples (Table S1, Fig. S1).

Raw transcriptomics data were downloaded when available and later processed using the recommended method for each microarray platform (i.e., Affymetrix (Santa Clara, CA, USA), Agilent (Santa Clara, CA, USA), and Illumina (San Diego, CA, USA)).

2.1.1. Affymetrix platforms data processing

Raw expression data from two-color Affymetrix platforms (Table S1) were processed using robust multiarray average algorithm (RMA) implemented in the *AFFY* package version 1.56 available through the *BIOCONDUCTOR* software project (<https://bioconductor.org>).

Probeset-to-gene mapping was done using BioMart web services via *BIOMART R* package version 2.34 [18], selecting the most expressed probe as representative of gene expression when multiple mapping probes occurred to avoid duplicated genes.

2.1.2. Two-color Agilent and CHUGAI platforms data processing

Raw expression data from two-color Agilent and CHUGAI platforms (Table S1) were processed using

minimum background correction method as implemented in the *backgroundCorrection* function of the LIMMA package available in R (<https://www.r-project.org/>). Background correction accounts for possible biases related to non-specific binding or spatial heterogeneity across the array. The next step in the normalization process is correcting for dye biases due to the presence of two colors in the array. This correction was performed using the loess method from the *normalizeWithinArrays* function also included in the LIMMA package. This method returns a matrix of corrected M and A values using the following expressions:

$$M = \log_2(R/G) = \log_2(R) - \log_2(G),$$

$$A = \frac{1}{2} \log_2(RG) = \frac{1}{2} (\log_2(R) + \log_2(G)).$$

The idea is to scale the log-ratios to have the same median absolute deviation (MAD) across samples. After normalizing each sample for dye biases, a normalization step between samples is needed to make them comparable with each other. This is achieved using the quantile method of the *normalizeBetweenArrays* function within the R LIMMA package. Finally, the normalized intensity values for the sample channel (i.e., red or green depending on the array design) are retrieved by solving the above-mentioned expressions, using the already calculated and normalized M and A values.

The probe-to-gene annotation was performed using the R package BIOMART version 2.34. When multiple probes mapped to the same gene, the most expressed one was selected to obtain a single representative probe for each gene. Then, HGNCHELPER package was used for the identification and correction of obsolete or invalid gene symbols to harmonize all datasets.

2.1.3. Illumina Beadchip Platforms data processing

Raw expression data from Illumina BeadChip Platforms (Table S1) were processed using the RMA background correction method as implemented in the *backgroundCorrection* function of the LIMMA package. Secondly, since this is a single-channel platform there is no need to perform a within-sample normalization, although between-sample normalization is still required. In this case, this is achieved using the quantile normalization method of the *normalizeBetweenArrays* function within the LIMMA package. The quantile approach makes the distribution of microarray signals the same between all arrays, making samples comparable between them. Then, HGNCHELPER package was used for the identification and correction of obsolete or invalid gene symbols to harmonize all datasets.

The probeset-to-gene annotation was performed using the R package BIOMART version 2.34. When multiple probes mapped to the same gene, the most expressed one was selected to obtain a single representative probe for each gene. Then, HGNCHELPER package was used for the identification and correction of obsolete or invalid gene symbols to harmonize all datasets.

For the case of TCGA-LUAD RNA-seq dataset, transcripts per million processed data were downloaded from TCGA2BED FTP repository [19].

2.2. LUAD consensus pathway transcriptional subtype definition framework

LUAD consensus transcriptional subtype classification framework is depicted in Fig. S2. Briefly, Gene Set Variation Analysis (GSVA) algorithm was used to evaluate the activity level of the 50 pathways included in the MSigDB hallmarks collection in each dataset, using a k -fold approach ($k = 5$) across 100 iterations [9,10]. Uniform Manifold Approximation and Projection (UMAP) dimension reduction method and walktrap clustering (Euclidean distance) were subsequently conducted on the previously obtained GSVA scores matrices to identify potential LUAD subpopulations [20]. Summary metrics for each potential LUAD subpopulation were calculated and used to establish final LUAD consensus subtypes using UMAP and walktrap method. Finally, tumor samples were assigned to the subtype to which they had been assigned the majority of times across the classification framework.

2.3. LUAD molecular subtype characterization

2.3.1. Clinicopathological covariates and overall survival

Association with clinicopathological variables (e.g., age, sex, stage, smoking status, and presence/absence genomic alterations) was assessed using COMPAREGROUPS package for R (V.4.2.0) [21]. Data regarding the presence/absence of LUAD oncogenic alterations (e.g., *EGFR*, *KRAS*, *ALK*, *TP53*, and *STK11*) were collected from the clinical data of the datasets included in this study when available.

The Cox proportional hazards models adjusted for age, sex, stage, smoking status, and study were used to test for the impact of our classification on overall survival (OS) rate.

2.3.2. Genomic characterization

TCGA-LUAD dataset [18] had available somatic alterations data for evaluating tumor mutational burden

(TMB) and COSMIC v3 mutational signatures [22]. For TMB, the total number of alterations per sample was assessed excluding synonymous variants. These values were then divided by the number of megabases (Mb) covered by the TCGA-LUAD whole-exome sequencing (WES) panel to obtain the number of mutations per Mb or TMB. Using somatic single nucleotide variants (SNV), mutational signatures were inferred using the R package SIGPROFILEREXTRACTOR [23].

Copy number alteration (CNA) levels were also evaluated in the TCGA-LUAD dataset [18]. Finally, genome instability was assessed using previously calculated DNA damage repair (DDR) deficiency scores in the TCGA-LUAD dataset [24].

2.3.3. Impact of the LUAD molecular classification on the immune landscape and immunotherapy response

The immune infiltrate composition of each LUAD sample was inferred using GSVA algorithm [10]. Gene signatures of the 21 evaluated immune fractions were obtained from a previous study [25]. Due to GSVA methodological constraints, single-gene signatures were replaced by their multi-gene counterparts published in a different study [26]. In addition, we also used specific cell categories when available, instead of the more generic supercategory (i.e., *M1 macrophages* and *M2 macrophages* instead of the broader *macrophages* category). For each cell type, we calculated the percentage of enriched tumors. Median GSVA scores for each cell fraction were used as the cut-off to define whether a sample is enriched in a specific cell type.

The status of a set of immune checkpoint inhibitors (ICI), activators (ICA), and T-cell effector and exhaustion markers was also evaluated [27,28]. For each gene expression dataset, the median gene expression value of each marker was used as the cut-off point for deciding whether a sample is enriched for a specific immune biomarker.

The predicted response to immunotherapy treatment was derived from the Tumor Immune Dysfunction and Exclusion (TIDE) scores already calculated for TCGA-LUAD dataset [29]. TIDE-positive scores indicate that a sample is less likely to respond to immunotherapy, because of the presence of immunosuppressive signals, whereas negative scores indicate potential response to immune checkpoint treatment (i.e., anti-CTLA4 and anti-PD-1). Binomial generalized linear models adjusted for PD-L1 gene expression and TMB values were used to test the impact of our classification on potential immunotherapy response.

2.4. Consensus transcriptional subtype independent validation

Subtyping of new samples in the CPTAC-3 validation cohort was inferred using the *predict* function of the *umap R* package version 0.2.7.0 and a *k*-nearest-neighbors approximation [20,30]. In summary, for each sample we obtained GSVA scores of the same 50 molecular pathways used to establish the original classification of LUAD tumors. This step was performed following the same steps previously described for the LUAD consensus pathway transcriptional subtype definition (fivefold, 100 iterations). Then, for each iteration, these GSVA scores were passed as an input to the *predict* function that produces 2D coordinates to map new samples onto the consensus map of LUAD tumors. New samples' subtype was predicted based on the most frequent label of the closest neighbors in the original classification. Therefore, after 100 iterations, each validation sample had 100 putative group assignments. Finally, samples were allocated to the AD subtype to which they had been assigned the majority of times throughout the classification process.

2.5. Identification of potential therapeutic vulnerabilities

Drug sensitivity data from three large pharmacogenomics studies were integrated to identify potential therapeutic vulnerabilities for each subtype using PHARMACO GX BIOCONDUCTOR/R package [31]. First, LUAD cancer cell lines (LUAD-CCL) were classified based on the primary tumor's classification using the *predict* function within *umap R* package as previously described for the cancer cell lines classification. Area above the curve (AAC) sensitivity measures for each drug and cell line were used to identify potential therapeutic vulnerabilities for the different subtypes. Importantly, *PharmacoGx* AAC values were normalized by the concentration range of the experiment in each study and take values between [0, 1]. Thus, the greater the AAC the more effective is a drug against a specific cell line. Subtypes were considered as potentially sensitive to the treatment if the average AAC value for the cell lines classified within a certain group was greater than the mean AAC plus 2 standard deviations for the drugs assessed in at least 2 out of the 3 pharmacogenomics studies. Also, average AACs were only calculated if the treatment had been tested in at least 2 different cell lines within a subtype and study.

3. Results

3.1. Consensus classification based on expression of 50 landmark molecular pathways yielded seven transcriptional LUAD subtypes

GSVA was conducted on more than 4500 LUAD in order to establish a consensus transcriptional classification based on the activity levels of 50 signaling pathways (see Section 2; Table S1) [9,10]. Using this approach, we identified seven LUAD transcriptional-based subtypes, labeled as AD1-7 (Fig. 1A). These subtypes were not evenly distributed throughout the whole set of tumors analyzed in this study (Fig. 2B). The most represented subtype was AD5 accounting for 20.95% of the tumors, whereas AD7 represented only 2.86% of the tumors.

Based on the relative activity of the signaling molecular pathways, each group displayed a specific transcriptional fingerprint (Fig. 1C, Fig. S3). A summary of the relatively upregulated and downregulated pathways within each LUAD subtype is depicted in Table 1.

3.2. LUAD transcriptional subtypes are correlated with clinicopathological covariates, distinct genomic profile, and overall survival

We evaluated the correlation of these subgroups with clinicopathological characteristics and whether they are represented across all the datasets included in the study (Table 2, Table S2). We observed a significant association with all evaluated covariates. Subtypes were represented in the different studies, although some subtypes may be more represented and underrepresented in certain datasets, most likely due to intrinsic biases of retrospective studies.

We also evaluated the association of the LUAD subtypes with the presence of clinically relevant driver oncogenic alterations (Table 2). *EGFR* mutations occurred more frequently in AD1, AD4, and AD7 groups, while *TP53* mutations were more common in AD2 and AD6 subtypes, and *STK11* alterations were enriched in AD1 and AD2 subtypes. *KRAS* mutations and *ALK* rearrangements were not correlated with any of the subgroups.

We also assessed whether this classification was associated with overall survival (OS) (Fig. 2, Fig. S4). AD1, AD4, and AD5 patients were associated with longer OS whereas, overall, AD2, AD3, AD6, and AD7 showed worse survival outcomes. This analysis was adjusted for the following covariates: age, gender, tumor stage, smoking history, and dataset.

3.3. LUAD pathway transcriptional profiling-based subtypes further subdivide previous mRNA-based subtypes

We performed a comparison with the previous LUAD mRNA-based consensus classification (bronchioid, squamoid, and magnoid) first described by Hayes et al. and later adopted by Wilkerson et al. and the TCGA for further exploration [18,32,33] (Fig. S5). Bronchioid mRNA subtype better aligned with AD1, AD4, and AD5 subtypes, all of them showing lower expression of proliferation-related pathways (Fig. 1C). AD1, AD4, and AD5 had consistently better OS as described for bronchioid tumors, when compared to squamoid or magnoid subtypes. Also, bronchioid subtype was enriched for *EGFR* mutations, which was also observed in AD1 and AD4 subtypes. Squamoid mRNA subtypes were for the most part associated with AD3, AD5, and AD6 subtypes, all of them showing higher expression of immune-related functions (Fig. 1C). This correlates with the higher immune cells infiltration previously found for squamoid tumors [34]. Moreover, AD6 was found to be enriched in *TP53* mutations; a trait also described for squamoid mRNA subtype. Finally, magnoid subtype mainly overlapped with AD2 proliferative subtype, which was enriched for *TP53/STK11* mutations (Fig. 1C, Table 2). Overall, these results demonstrate concordance among both LUAD classifications, but previous mRNA-based subtypes were further subdivided by using our approach.

3.4. LUAD transcriptional subtypes were also correlated with tumor mutational burden and DNA damage

Using the TCGA-LUAD dataset, LUAD subtypes were further characterized at the genomic level [18]. First, using whole-exome sequencing data we evaluated potential differences in terms of TMB and mutational signatures included in the COSMIC v3 collection (Fig. 3A,B) [22]. TMB significantly differed among LUAD transcriptional subtypes (Fig. 3A). AD2 and AD6, which are also enriched for *TP53* mutations, had significantly higher TMB values when compared to the rest of subtypes, except for AD7 (Table S3). Concerning COSMIC mutational signatures, tobacco and clock-like signatures were overrepresented across LUAD subtypes (Fig. 3B). Notably, our results showed a significant association between the subtypes and the prevalence of mutational signatures SBS1 (clock-like), SBS4 (tobacco), and SBS13 (APOBEC activity) (Table S4).

Fig. 1. LUAD subtype pathway transcriptional landscape. (A) LUAD consensus map of pathway transcriptional profiling-based subpopulations. Each dot represents the summary centroid of the different subpopulations identified during the classification process. Using UMAP and walktrap clustering method with Euclidean distance on these centroids, seven different consensus groups, represented by different colors, were identified based on the joint behavior of the 50 studied molecular pathways. (B) Barplots representing the distribution of LUAD tumors across the seven transcriptional subtypes. (C) Heatmap representing relative activity levels (GSVA scores) of the 50 studied pathways (rows) in each of the 4573 LUAD tumor samples (columns) that were assigned to a consensus subtype. Red colors indicate higher relative activity of a pathway in a certain sample, whereas blue colors indicate lower relative activity of a pathway in a certain sample.

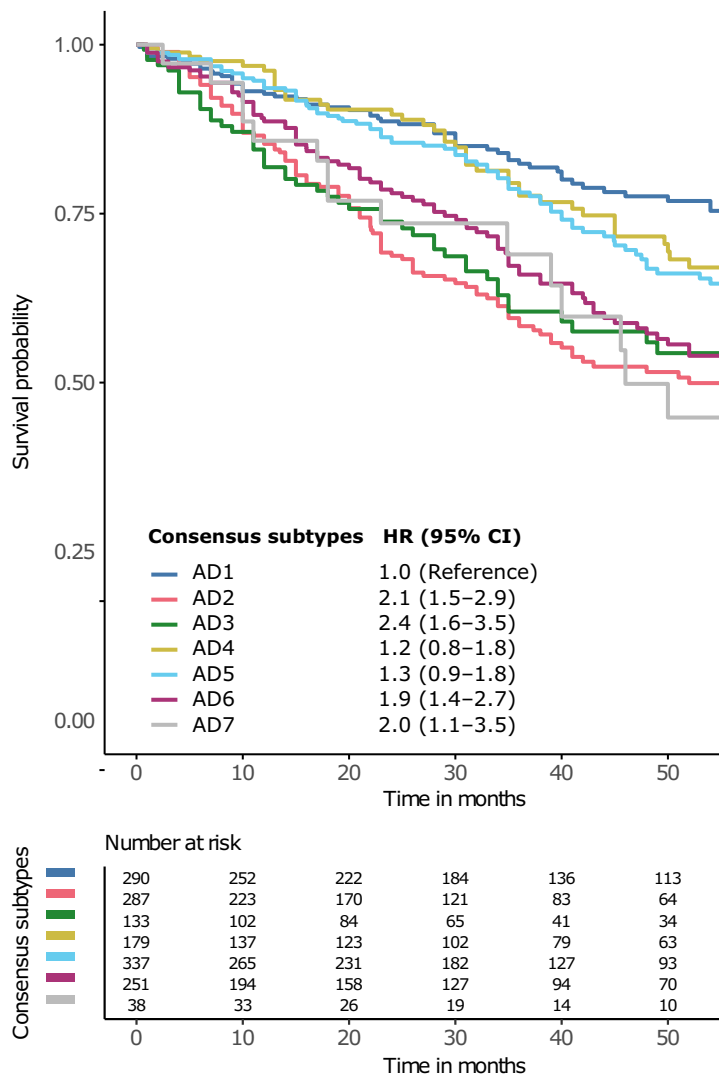


Fig. 2. Overall survival by LUAD subtype. Kaplan–Meier curves of each of the identified pathway transcriptional profiling-based LUAD groups. Hazard ratios (HR) and 95% confidence intervals (95% CI) come from a Cox proportional hazards model adjusted for age, sex, stage, smoking history, and dataset. For this analysis, we used the subset of datasets with available survival data and complete covariates information for the Cox proportional hazards model ($n = 10$ datasets, $n = 1515$ samples).

Using TCGA-LUAD data [18], copy number alterations (i.e., amplifications or deletions) were more common in AD2 and AD6 subtypes compared to the rest of the subtypes, except for AD7 (Fig. 3C, Table S5). We also assessed the level of genomic

instability and DNA damage repair (DDR) capacity according to these subgroups (Fig. 3D). Again, AD2 and AD6 samples showed significantly higher DDR deficiency scores than the rest of subtypes (Table S6).

Table 1. Molecular pathway landscape across LUAD subtypes.

Consensus subtype	Upregulated pathways	Downregulated pathways
AD1	Metabolic pathways	Angiogenesis Epithelial–mesenchymal transition Immune-related pathways Cell cycle-related pathways PI3K-AKT–MTOR signaling
AD2	DNA repair Oxidative phosphorylation Cell cycle-related pathways	Angiogenesis Epithelial–mesenchymal transition Immune-related pathways Apoptosis TGF-B signaling Hedgehog/Notch signaling IL2-STAT5 signaling
AD3	Angiogenesis Epithelial–mesenchymal transition Immune-related pathways Metabolic pathways Apoptosis Hypoxia Protein secretion TP53 pathway KRAS signaling IL2-STAT5 signaling TNFA via NFkB signaling PI3K-AKT–MTOR signaling TGF-B signaling Notch signaling MTORC1 signaling	
AD4		DNA repair Metabolic pathways Apoptosis Hypoxia Protein secretion Cell cycle-related pathways KRAS signaling PI3K-AKT–MTOR signaling TGF-B signaling MTORC1 signaling Unfolded protein response
AD5	Immune system-related pathways Apoptosis TP53 pathway IL2-STAT5 signaling TNFA via NFkB signaling Hedgehog signaling	DNA repair Metabolic pathways Cell cycle-related pathways Unfolded protein response
AD6	DNA repair Interferon-gamma Interferon alpha Metabolic pathways Cell cycle-related pathways PI3K-AKT–MTOR signaling MTORC1 signaling Unfolded protein response	Hedgehog signaling WNT B-catenin signaling
AD7	Estrogen response Notch signaling	

Table 2. Correlation of clinicopathological variables with LUAD subtypes. The number of samples with available information in each case is depicted in the *N* column. MUT, mutated; WT, wildtype.

	<i>N</i>	AD1 (%) <i>N</i> = 766	AD2 (%) <i>N</i> = 851	AD3 (%) <i>N</i> = 473	AD4 (%) <i>N</i> = 598	AD5 (%) <i>N</i> = 952	AD6 (%) <i>N</i> = 774	AD7 (%) <i>N</i> = 129	<i>P</i>
Sex, <i>N</i> (%)	3906								< 0.001
M		324 (50.47)	427 (56.86)	217 (51.91)	223 (42.72)	356 (43.41)	331 (50.46)	48 (49.48)	
F		318 (49.53)	324 (43.14)	201 (48.09)	299 (57.28)	464 (56.59)	325 (49.54)	49 (50.52)	
Age, <i>N</i> (%)	3609								< 0.001
≤ 50		55 (9.18)	77 (11.16)	33 (8.62)	34 (7.02)	61 (8.07)	61 (10.13)	14 (14.74)	
> 50–65		245 (40.90)	325 (47.10)	148 (38.64)	198 (40.91)	282 (37.30)	271 (45.02)	41 (43.16)	
> 65		299 (49.92)	288 (41.74)	202 (52.74)	252 (52.07)	413 (54.63)	270 (44.85)	40 (42.11)	
Stage, <i>N</i> (%)	3128								0.005
Early-stage (I–II)		472 (86.61)	493 (80.42)	260 (81.00)	319 (85.07)	578 (85.50)	408 (79.53)	67 (78.82)	
Late-stage (III–IV)		73 (13.39)	120 (19.58)	61 (19.00)	56 (14.93)	98 (14.50)	105 (20.47)	18 (21.18)	
Smoking history, <i>N</i> (%)	2788								< 0.001
Never smoker		152 (30.83)	64 (12.19)	56 (20.29)	99 (26.83)	164 (27.89)	81 (17.16)	16 (24.62)	
Smoker		341 (69.17)	461 (87.81)	220 (79.71)	270 (73.17)	424 (72.11)	391 (82.84)	49 (75.38)	
<i>EGFR</i> mutation, <i>N</i> (%)	1537								< 0.001
WT		185 (62.93)	231 (79.66)	105 (76.09)	111 (63.43)	248 (71.06)	195 (77.38)	24 (61.54)	
MUT		109 (37.07)	59 (20.34)	33 (23.91)	64 (36.57)	101 (28.94)	57 (22.62)	15 (38.46)	
<i>KRAS</i> mutation, <i>N</i> (%)	1360								0.239
WT		184 (72.73)	173 (70.90)	79 (63.71)	124 (78.48)	225 (72.12)	169 (73.48)	28 (71.79)	
MUT		69 (27.27)	71 (29.10)	45 (36.29)	34 (21.52)	87 (27.88)	61 (26.52)	11 (28.21)	
<i>ALK</i> translocation, <i>N</i> (%)	456								0.064
WT		99 (94.29)	67 (83.75)	24 (85.71)	52 (85.25)	83 (91.21)	76 (96.20)	10 (83.33)	
MUT		6 (5.71)	13 (16.25)	4 (14.29)	9 (14.75)	8 (8.79)	3 (3.80)	2 (16.67)	
<i>TP53</i> mutation, <i>N</i> (%)	849								< 0.001
WT		128 (85.91)	83 (50.30)	65 (78.31)	69 (75.00)	171 (85.93)	80 (56.34)	14 (73.68)	
MUT		21 (14.09)	82 (49.70)	18 (21.69)	23 (25.00)	28 (14.07)	62 (43.66)	5 (26.32)	
<i>STK11</i> mutation, <i>N</i> (%)	598								< 0.001
WT		80 (75.47)	87 (73.11)	53 (92.98)	63 (92.65)	128 (91.43)	87 (91.58)	12 (92.31)	
MUT		26 (24.53)	32 (26.89)	4 (7.02)	5 (7.35)	12 (8.57)	8 (8.42)	1 (7.69)	

3.5. LUAD molecular subtypes had distinct immune cells infiltration patterns and were associated with different immunotherapy responses

The immune infiltrate composition of each sample was quantified by applying GSVA on 21 immune cell-specific gene signatures (Fig. 4A, Fig. S6) [25]. On the one hand, AD3 and AD5 tumors displayed higher infiltration of most immune cells, including both immune active and immunosuppressive categories. Nevertheless, there were also some distinctive features between AD3 and AD5 LUAD subtypes. For instance, AD3 subtype comprised a higher percentage of tumors with high Th2 infiltration when compared to AD5. AD4 subtype was preferentially infiltrated by innate immune cells (i.e., NK cells, neutrophils, eosinophils, and mast cells) and some specific T-cell populations (i.e., follicular T helper cells, T effector memory cells, T effector memory cells, T gamma-delta cells, and T

helper 17). However, in AD4 subtype the presence of immunosuppressive cells (i.e., macrophages M2 and T regulatory cells) was lower than in other highly infiltrated subtypes (i.e., AD3, AD5, and AD6). AD6 tumors appeared to be more frequently enriched by T-cell populations, both with cytotoxic and immunosuppressive roles (i.e., cytotoxic T cells, T regulatory, T helper 1, and T helper 2). AD6 subtype was also commonly infiltrated by other immunosuppressive cells (i.e., macrophages) and other innate cells (i.e., active dendritic cells and CD56dim NK cells). Finally, AD2 was, overall, the least infiltrated subtype compatible with an immune desert phenotype.

Regarding immune checkpoint and T-cell expression markers, AD3, AD5, and AD6 were also enriched in tumors showing higher expression levels of a wide variety of the evaluated biomarkers, followed by AD4 (Fig. 4B, Fig. S7).

Finally, we also evaluated the utility of our LUAD subtypes to predict the predisposition to immunotherapy

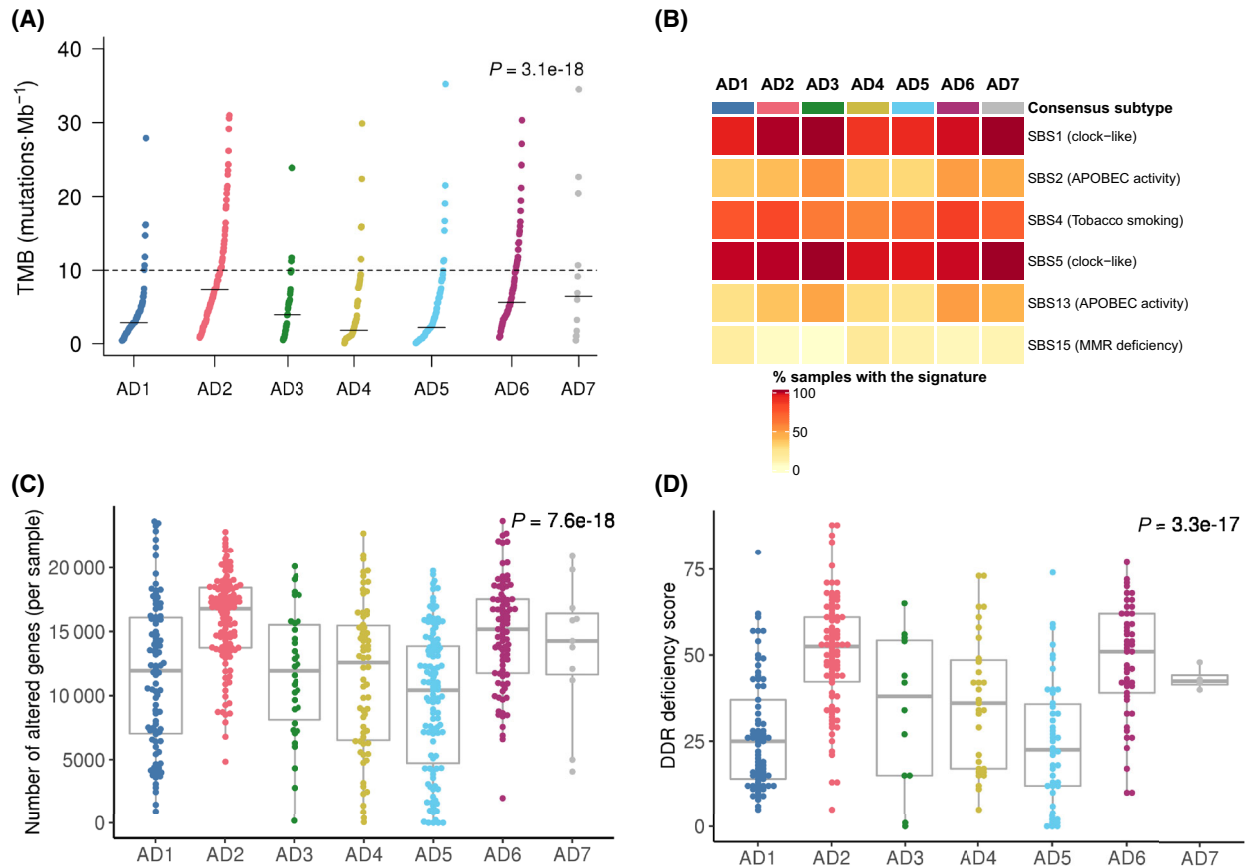
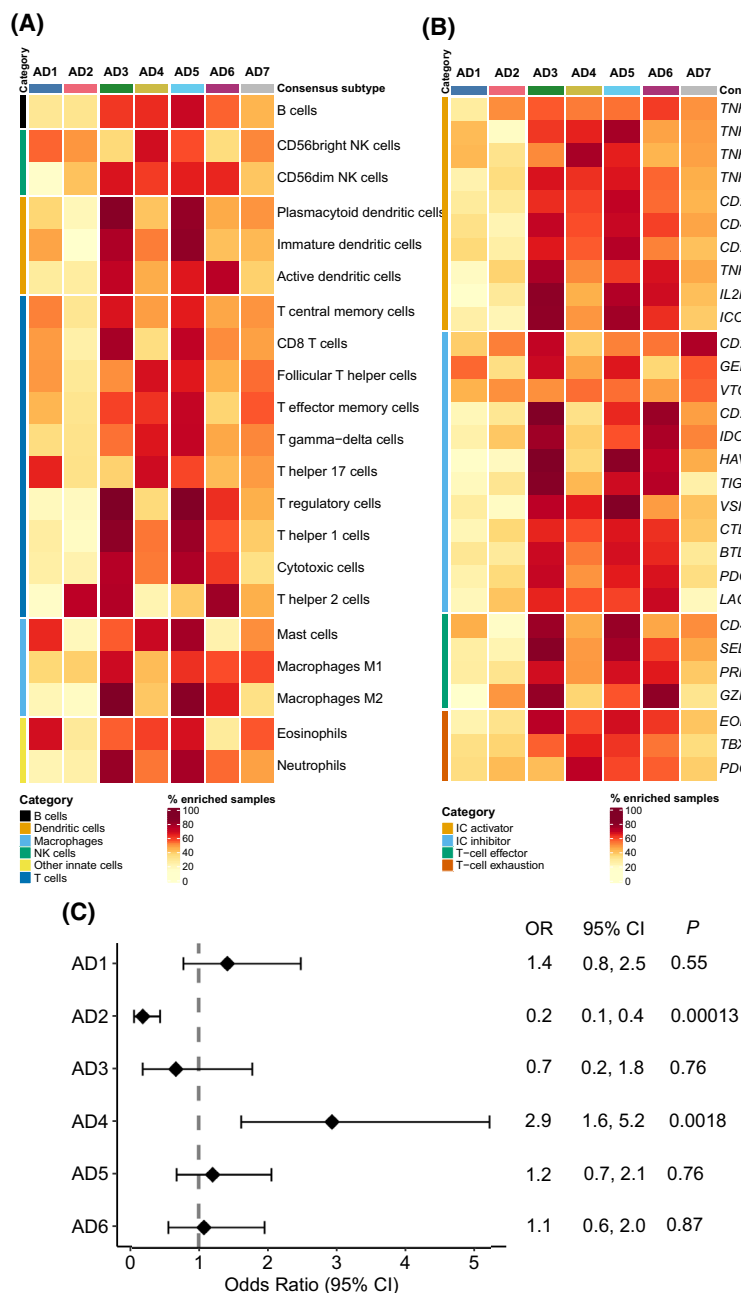


Fig. 3. Genomic characterization in the TCGA-LUAD set. (A) Tumor mutational burden (TMB) across LUAD consensus subtypes. Each dot represents the TMB value for a specific sample. The black segment represents the median TMB value for each LUAD subtype. The horizontal dotted line represents 10 mutations-Mb⁻¹ TMB value, which is a common cut-off for designating TMB high or low. Kruskal-Wallis test was used to assess potential differences regarding TMB between LUAD subtypes. P value was corrected using the false discovery rate (FDR) multiple-comparison correction method. (B) Heatmap representing the percentage of positive samples for each specific COSMIC mutational signature (rows) in each LUAD subtype (columns). Samples were designated as positive if they harbored at least one mutation associated with a certain mutational signature. (C) Boxplots of the copy number alterations burden across LUAD subtypes. Each dot represents the number of altered genes per sample. Kruskal-Wallis test was used to assess potential differences regarding the number of copy number altered genes between LUAD subtypes. P value was corrected using the false discovery rate (FDR) multiple-comparison correction method. (D) Boxplots of the DNA damage repair (DDR) deficiency score distribution across LUAD subgroups. Each dot represents the DDR score per sample. Kruskal-Wallis test was used to assess potential differences regarding DDR scores between LUAD subtypes. P value was corrected using the false discovery rate (FDR) multiple-comparison correction method. For A–D, only the TCGA-LUAD dataset was used as it is the only one with associated transcriptomics and genomics data. Number of samples of each subtype are: AD1: 90, AD2: 114, AD3: 34, AD4: 66, AD5: 113, AD6: 85, AD7: 12.

response beyond PD-L1 and TMB biomarkers using previously calculated TIDE scores in the TCGA-LUAD dataset [29]. We used a likelihood ratio test to compare two binomial generalized linear models (GLM) predicting immunotherapy response (i.e., yes or no). The first GLM included PD-L1 gene expression (i.e., low and high based on median cut-off) and TMB values as independent variables, and the second GLM was identical but also considering LUAD subtype as a predictor. Results showed that LUAD subtype further contributes to

predict the probability of immunotherapy response ($P = 0.0003$). Moreover, and although not used as a stratification criterion in NSCLC in clinical trials or in the clinical practice, we also added PD-1 expression (i.e., low and high based on median cut-off) as a proxy of T-cell infiltration to the model. Again, the results showed that our classification further contributes to predict the probability of immunotherapy response ($P < 0.001$). Given this outcome, for each subtype, we assessed the likelihood of immunotherapy response when compared

Fig. 4. Immune characterization and association with immunotherapy response. (A) Heatmap representing the percentage of samples showing high relative infiltration of 21 evaluated immune cell types. Median immune cell abundance GSEA score values were used as a cut-off to designate if a sample is enriched for a specific immune cell. Different immune cell categories are represented with different colors on the left side of the heatmap. (B) Heatmap representing the percentage of samples of samples with high expression of a set of immune-related biomarkers. Median gene expression values for each gene in each gene expression dataset were used as a cut-off to designate if a sample is enriched for a specific biomarker. Different immune marker categories are represented with different colors on the left side of the heatmap. (C) Forest plot showing the odds ratios, confidence intervals, and FDR-adjusted *P* value for immunotherapy response in each LUAD subtype when compared to all other subtypes. Odds ratio for AD7 subtype could not be calculated as 0 patients were predicted as potential responders in this subtype. A and B analysis were performed considering all gene expression datasets ($n = 4573$ LUAD samples). For C, we used pre-computed TIDE scores for the TCGA-LUAD dataset ($n = 486$, AD1: 86, AD2: 105, AD3: 32, AD4: 64, AD5: 106, AD6: 81, AD7: 12).



to the tumors in any other subtypes (Fig. 4C). Tumors within AD4 subtype were found to be 2.9 times more likely to respond to immunotherapy compared to the tumors classified in any other subtype (34.4% predicted responders in AD4 [$n = 64$] vs 15.2% in other subtypes [$n = 422$]). Despite being among the most infiltrated subtypes and showing high PD-L1 gene expression (Fig. 4A, B), only 12.5% of AD3 tumors were predicted as potential responders. Also, in correlation with its immune excluded phenotype, AD2 tumors were 80% less likely to

respond to immunotherapy than other subtypes (4.76% predicted responders in AD2 [$n = 105$] vs 21.2% in other subtypes [$n = 381$]).

3.6. LUAD consensus subtype independent validation

We conducted an independent validation of the LUAD subtypes using CPTAC-3 LUAD dataset [30]. The activity level of 50 molecular pathways was

measured and mapped in 111 LUAD which were classified based on a k -nearest-neighbors algorithm (Fig. 5A). All seven subtypes were predicted in this independent dataset, confirming the robustness of the classification. Moreover, the pathway transcriptional footprint of each subtype is conserved between the original and the validation datasets (Fig. S8). To further prove the validity of the predictions, we explored whether the association between the LUAD subtype and copy number alterations, and TMB is conserved in the validation set (Fig. 5B,C). Notably, subtype TMB and copy number alterations rate are highly concordant between the original and validation sets, confirming that previously found associations at the genomic level are maintained (Figs 3A,C and 5B,C).

3.7. Analysis of drug sensitivity in *in vitro* data revealed potential therapeutic vulnerabilities for the subtypes

Data from three large-scale pharmacogenomics studies conducted on cancer cell lines were integrated to explore potential therapeutic vulnerabilities in LUAD. First, LUAD-CCLs were classified according to the primary tumors' classification, and then, we assessed the impact of our classification on the response to specific compounds (Fig. 6, Fig. S9).

LUAD-CCL subtypes were considered potentially sensitive to a specific drug whenever average AAC values were greater than the mean plus 2 standard deviations of all drugs AAC values in at least 2 out of the 3 evaluated studies. Out of 239 evaluated drugs (i.e., number of drugs tested in at least two studies), only 5 were found to be consistently effective (i.e., AAC values above threshold) in at least two studies for some of the subtypes and not the others. Overall, cells assigned to AD2 showed potential sensitivity to vincristine and gemcitabine chemotherapies, which correlates with its proliferative nature. Also, cell lines classified in AD3, AD6, and AD7 subtypes, also showing high cell cycle activity, were found to be potentially sensitive to gemcitabine treatment. Interestingly, AZD7762 CHK1 inhibitor could be potentially suitable for AD2 cell lines, which correlates with the higher genome instability described for AD2 subtype. Despite a lower cycling nature of subtype AD1 and AD4, cell lines classified within these subtypes appeared to be potentially sensitive to dinaciclib, based on these data.

4. Discussion

In this study, we integrated the transcriptional profiles of more than 4500 LUAD, and based on the activity

levels of a set of 50 molecular pathways, we were able to identify seven LUAD molecular subtypes. Importantly, the number of samples included in this study further exceeds that of previous studies, covering the largest part of the molecular diversity of LUAD [8]. This classification was associated with survival outcomes and was correlated with relevant clinical characteristics. Besides, at the genomic level, LUAD transcriptional subtypes were associated with the presence of oncogenic driver alterations, mutational signatures, CNA burden, and DDR capacity. These results support the previously described transcriptional heterogeneity that exists within LUAD histological entity [8]. Furthermore, the integration of drug sensitivity data from three large pharmacogenomics studies unraveled potential therapeutic vulnerabilities for the subtypes. Finally, the transcriptional subtypes showed distinct patterns in terms of immune cells infiltration and immune-related biomarkers expression and were able to predict immune response in addition to PD-L1 gene expression and TMB.

Since early 2000s, there have been several efforts to define clinically relevant LUAD transcriptional subtypes, which resulted in various different classifications [8,30,35,36]. Despite all these studies, LUAD subtypes have never been translated into the clinical setting. Reasons for this include intrinsic technical and analytical limitations, such as low overlap between the gene signatures, probably due to intrinsic technical and biological variability of individual gene expression levels. In our work, we focused on the activity levels of a set of established molecular pathways rather than in the expression of individual genes. This approach is likely to reduce the effect of the stochastic sources of variability to which multiple single-gene measures are subjected [9]. Moreover, the method used for measuring the pathway activity (GSVA algorithm) is able to overcome batch effects compared with other deconvolution methods [10,37]. Importantly, we were able to validate our classification framework in an independent set of samples [30]. This approach would therefore be capable to accurately classify new prospective samples into one of the specific transcriptional subtypes.

Also, we evaluated the correspondence between the widely accepted Hayes et al. mRNA subtypes and the present classification [21,32,33]. In summary, we found that, in most cases, our pathway transcriptional profiling-based subtypes further stratified the ones proposed by Hayes et al., based on individual genes expression, suggesting a higher resolution of our classification to deal with the molecular heterogeneity that exists within LUAD.

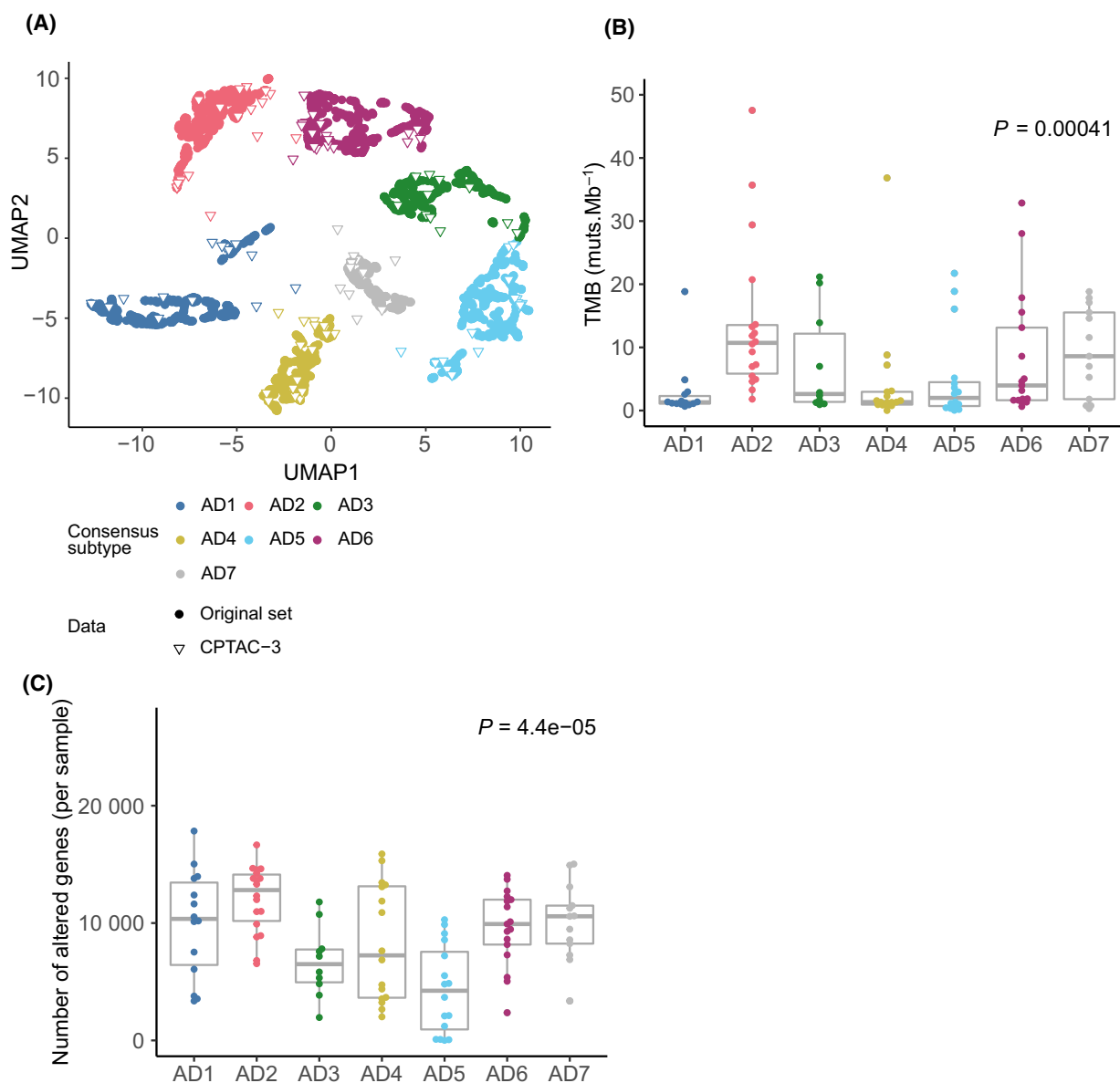


Fig. 5. LUAD pathway transcriptional profiling-based classification independent validation. (A) New samples from the CPTAC-3 LUAD dataset were mapped on the previously established LUAD classification. New samples' subtype status was decided based on the most frequent label of the 51 nearest neighbors of the original classification. Colored circles represent samples used in the original set, whereas triangles represent new CPTAC-3 validation set samples ($n = 105$). (B) Boxplot representing tumor mutational burden (TMB) values across newly classified CPTAC-3 LUAD samples. Each dot represents the TMB value per sample (AD1:14, AD2: 18, AD3: 10, AD4: 17, AD5: 16, AD6: 17, AD7: 13). Kruskal-Wallis test was used to assess potential differences regarding TMB between LUAD subtypes. P value was corrected using the false discovery rate (FDR) multiple-comparison correction method. (C) Boxplot representing copy number burden across newly classified CPTAC-3 LUAD samples. Each dot represents the number of altered genes per sample (AD1:14, AD2: 18, AD3: 10, AD4: 17, AD5: 16, AD6: 17, AD7: 13). Kruskal-Wallis test was used to assess potential differences regarding the number of copy number altered genes between LUAD subtypes. P value was corrected using the false discovery rate (FDR) multiple-comparison correction method.

The lack of association of previously described LUAD intrinsic subtypes with available therapeutic strategies prevented their clinical use. We tried to

overcome this limitation by integrating drug sensitivity data from *in vitro* pharmacogenomics studies [31]. These databases have greater drug coverage compared

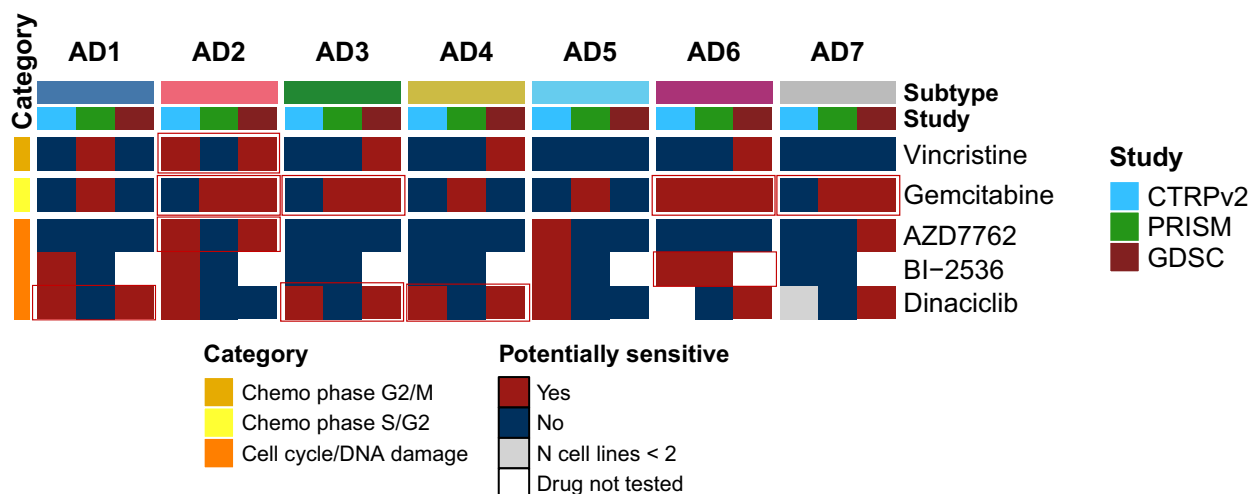


Fig. 6. Potential treatment strategies for the LUAD subtypes based on CTRPv2, GDSC, and PRISM LUAD-CCLs drug sensitivity data. Heatmap representing potentially suitable specific therapeutic strategies in at least two different studies within the same subtype. Mean area above the curve (AAC) sensitivity metric values were only calculated if the drug had been tested in at least two different LUAD-CCLs within a subtype and study. Subtypes were considered as potentially sensitive to the treatment if the average AAC value for the cell lines classified within a certain group was greater than mean AAC plus 2 standard deviations for the drugs assessed in at least 2 out of the 3 evaluated pharmacogenomics studies.

with other available ones such as CMap, which has been used for similar purposes [35]. Although significant discrepancies can exist between drug response results obtained from cancer cell lines and clinical response in patients, we were able to identify some potential drug candidates for the different LUAD subtypes, in line with their molecular characteristics. In this way, chemotherapy alone, or combined with immunotherapy, is the cornerstone for patients with driver-negative LUAD. However, clinical responses upon chemotherapy regimens are highly heterogeneous and underscore the need for improving patient selection [38]. In our work, we observed that AD2, AD3, AD6, and AD7 cell lines might benefit from vincristine and gemcitabine chemotherapies. Moreover, cancer cells classified as AD2 showed potential sensitivity to AZD7762 CHK1 inhibitor, which correlates with the higher genome instability seen in this subtype.

Genomic profiling is crucial in LUAD tumors to guide the most appropriate treatment based on the detection of actionable oncogenic alterations. In fact, this study does not intend to replace the current classification based on genomic profiling. However, there is a non-negligible percentage of patients lacking tractable genomic alterations, and even patients with oncogenic drivers show heterogeneous responses to targeted therapies for reasons that remain unclear, and all patients will eventually develop treatment resistance [39]. Our results highlight the significant heterogeneity

of this disease as patients with the same mutational event were found to be distributed across all subtypes, being *KRAS* mutant LUAD the most heterogeneous entity. In addition to the role of concurrent genomic alterations, differences in the activation of transcriptional pathways could explain that patients harboring identical driver alterations might have distinct clinical outcomes upon targeted therapy. Conversely, the fact that tumors with different driver alterations coexist in the same transcriptional subtype suggests that different oncogenic mutations may give rise to similar transcriptional phenotypes, which could benefit from similar combinatorial strategies. Therefore, the implementation of new methodologies beyond genomic testing, such as those based on gene expression, could help to deliver more precise and innovative treatments to patients with LUAD, specially in those patients without actionable genomic alterations or that have progressed frontline chemoimmunotherapy.

Immunotherapy alone or in combination with chemotherapy has become the standard of care for driver-negative metastatic LUAD [3]. However, durable clinical benefit is observed only in a reduced fraction of patients (< 20%) [40]. Previous studies have shown that TMB or PD-L1 expression cannot accurately predict long-term benefit in all patients [41]. The improvement of patient selection and the definition of rational combinations are therefore an unmet clinical need. Transcriptomic data could provide clinically

relevant information beyond individual markers. In this regard, our results showed that AD2, despite having high TMB was an immune cold subtype, and was 80% less likely to respond to immunotherapy than tumors classified in other subtypes. This result is concordant with the findings of a previous study that also identified a LUAD subtype with high TMB but no apparent immune infiltration [35]. Overall, these results underline the limitation of TMB to predict potential response to immunotherapy in LUAD [42]. We also found that although most patients classified in AD3 subtype showed higher cytotoxic T-cell infiltration and PD-L1 gene overexpression (*CD274*), they were also unlikely to respond to ICI therapy according to TIDE scores (12% patients were classified as responders in AD3) [43]. AD3 tumors not only co-express a wide variety of immune checkpoint inhibitors and T-cell exhaustion markers but also showed high infiltration of immunosuppressive cells, such as M2 macrophages and T regulatory cells, which could contribute to intrinsic resistance to immunotherapy. Thus, macrophage-targeted therapy could be a potential solution for improving AD3 tumor response [44]. Also, AD3 shows relatively high TGF- β signaling activity, which has previously been associated with lack of response to immunotherapy [45]. For this reason, rational combinations of ICI and immune cell-specific targeted therapies could probably improve clinical outcomes in solid tumors. However, most clinical trials are not yet selecting patients based on the immune contexture [46,47]. Tumors classified in AD4 subtype were 2.9 times more likely to respond to immunotherapy than tumors classified in other subtypes. These tumors showed infiltration of cytotoxic T cells and other cells involved in tumor destruction (i.e., B cells, NK cells, diverse types of T cells, etc.) and lower infiltration of immunosuppressive cells (e.g., T regulatory cells, macrophages M2, etc.), potentially constituting a less immune evasive microenvironment. Although further validation through other techniques that provide more cellular resolution (i.e., scRNA-seq) would be needed, these results underscore the need to comprehensively characterize the immune contexture, along with conventional single biomarkers (i.e., PD-L1 and TMB), to perform an accurate patient stratification and deliver tailored and effective treatment strategies for advanced LUAD.

Despite all the obtained results, our study has some intrinsic limitations that must be acknowledged. This is a retrospective analysis of multiple microarray and RNA-seq gene expression studies, which rely on fresh tissue biopsies. Thus, further research is needed towards the implementation of this classification in

formalin-fixed paraffin-embedded samples, which are routinely available in the clinical setting. For instance, we believe that with the incorporation of new profiling technologies, such as HTG EdgeSeq, which allows whole-transcriptome gene expression profiling in FFPE samples, it will be possible to evaluate the clinical relevance of our framework using clinical samples. Moreover, although later validated in the CPTAC-3 dataset, results regarding the association with TMB and CNA were based exclusively on the TCGA-LUAD dataset, as the rest of the studies did not have associated WES or CNA data. Most studies included patients who were surgically resected and did not receive systemic therapy, or this information was not available. For instance, this is particularly relevant for the results regarding immunotherapy response predictions, which should be further validated in retrospective and prospective studies of patients with LUAD treated with ICI. Regarding cancer cell lines drug sensitivity results, potential drug candidates are based on *in vitro* data and do not take into consideration the interplay between cancer cells and TME. However, these models are continuously used in preclinical research for similar purposes (i.e., drug screening and hypothesis generation) and we believe that this exercise could be useful to prioritize which compounds could be tested in more advanced preclinical models (i.e., tumoroids and patient-derived xenografts).

5. Conclusions

To sum up, we have presented and validated a robust and clinically relevant classification of LUAD tumors, based on the transcriptional activity levels of important cellular pathways. To our knowledge, no previous LUAD classification has been derived from such a large sample size. Despite significant challenges, we believe that the integration of transcriptomic and genomic data could improve patient stratification and may pave the way for guiding novel therapeutic approaches in patients with LUAD.

Acknowledgements

SH-P is supported by an AGAUR-FI fellowship (2022 FI_B2 00066), with the support of the FI program of the Secretariat for Universities and Research of the Department of Business and Knowledge of the Government of Catalonia, and the support of the European Union through the European Social Fund “ESF, Investing in your future.” XS received support from the Ministerio de Ciencia, Innovación y Universidades, which is part of the Agencia Estatal de Investigación

(AEI), through the Retos Research Grant, number RTI2018-102134-A-I00 (co-funded by the European Regional Development Fund, ERDF, a way to build Europe). EN received support from the Instituto de Salud Carlos III (grants PI18/00920 and PI21/00789) (co-funded by the European Regional Development Fund, ERDF, a way to build Europe). We thank the CERCA Programme/Generalitat de Catalunya for institutional support. RM is supported with the funding of the Ministerio de Universidades, through the predoctoral fellowship number FPU19/01734 for the Formación de Profesorado Universitario (FPU). NV and JB are supported by a Rio Hortega contract (CM19/00245 and CM21/00073, respectively) from the Instituto de Salud Carlos III. We sincerely thank Cristina Muñoz-Pinedo for useful comments that significantly helped to improve the present manuscript.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. XS participated in lectures from Roche. EN received research support from Roche, Pfizer, Merck Serono, and Bristol Myers Squibb and participated in advisory boards or lectures from Bristol Myers Squibb, Merck Serono, Merck Sharpe & Dohme, Lilly, Roche, Pfizer, Bayer, Sanofi, Takeda, Boehringer Ingelheim, Amgen, and AstraZeneca.

Author contributions

SH-P, XS, and EN contributed to conceptualization, formal analysis, data interpretation, writing – original draft, and writing – review and editing; DC, AA, RM, NV, RP, JB, and AM-C contributed to data interpretation, analysis, and writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Peer review

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/1878-0261.13550>.

Data accessibility

Gene expression data for LUAD consensus classification were obtained from GEO and ArrayExpress public repositories. Specific information and identifiers from each dataset are available at Table S1. CCLE and

GDSC LUAD cancer cell lines molecular data were obtained from <https://depmap.org> and https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html, respectively. Cellosaurus identifiers of specific cell lines used are depicted in Fig. S8. GDSC, CTRPv2, and PRISM studies drug sensitivity data were available within the PHARMACO GX R package [31]. LUAD CPTAC-3 study gene expression and copy number alterations data were downloaded from the supplementary material of [30].

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;**71**(3):209–49.
- Pirker R. Conquering lung cancer: current status and prospects for the future. *Pulmonology.* 2020;**26**(5):283–90.
- Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2018;**29**(Suppl 4): iv192–237.
- Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol.* 2009;**25**(4):195–203.
- Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med.* 2011;**17**(3):297–303.
- Majem M, Juan O, Insa A, Reguart N, Trigo JM, Carcereny E, et al. SEOM clinical guidelines for the treatment of non-small cell lung cancer (2018). *Clin Transl Oncol.* 2019;**21**(1):3–17.
- Anusewicz D, Orzechowska M, Bednarek AK. Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of notch, hedgehog, Wnt, and ErbB signalling. *Sci Rep.* 2020;**10**(1):21128.
- Hijazo-Pechero S, Alay A, Marín R, Vilariño N, Muñoz-Pinedo C, Villanueva A, et al. Gene expression profiling as a potential tool for precision oncology in non-small cell lung cancer. *Cancers (Basel).* 2021;**13**(19):4734.
- Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst.* 2015;**1**(6):417–25.
- Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics.* 2013;**14**(1):7.
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* 2012;**483**(7391):603–7.

- 12 Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 2013;**41** (Database issue):D955–61.
- 13 Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer.* 2020;**1**(2):235–48.
- 14 Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* 2015;**5**(11):1210–23.
- 15 ArrayExpress < EMBL-EBI [cited 2022 Jul 28]. Available from: <https://www.ebi.ac.uk/arrayexpress/>
- 16 Home – GEO – NCBI [cited 2022 Jul 28]. Available from: <https://www.ncbi.nlm.nih.gov/geo/>
- 17 Lung cancer explorer | lung cancer database | lung cancer analysis | Quantitative Biomedical Research Center | UT Southwestern [cited 2022 Nov 14]. Available from: <https://lce.biohpc.swmed.edu/lungcancer/>
- 18 Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;**511** (7511):543–50.
- 19 Cumbo F, Fisco G, Ceri S, Masseroli M, Weitschek E. TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas. *BMC Bioinformatics.* 2017;**18**(1):6.
- 20 Konopka T. umap: Uniform Manifold Approximation and Projection R package [cited 2023 Nov 15]. Available from: <https://github.com/tkonopka/umap>
- 21 Subirana I, Sanz H, Vila J. Building Bivariate Tables: The compareGroups Package for R. *Journal of Statistical Software.* 2014;**57**(12):1–16.
- 22 Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature.* 2020;**578**(7793):94–101.
- 23 Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics.* 2022;**2**(11).
- 24 Marquard AM, Eklund AC, Joshi T, Krzystanek M, Favero F, Wang ZC, et al. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark Res.* 2015;**3**:9.
- 25 Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity.* 2013;**39**(4):782–95.
- 26 Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 2017;**18** (1):248–62.
- 27 Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer.* 2012;**12** (4):252–64.
- 28 Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol.* 2015;**15** (8):486–99.
- 29 Tumor Immune Dysfunction and Exclusion (TIDE) [cited 2022 Sep 12]. Available from: <http://tide.dfci.harvard.edu/login/>
- 30 Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell.* 2020;**182**(1):200–225.e35.
- 31 Smirnov P, Safikhani Z, Eeles C, Freeman M, Haibe-Kains B. PharmacoGx: analysis of large-scale pharmacogenomic data. Bioconductor version: release (3.15).
- 32 Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol.* 2006;**24**(31):5079–90.
- 33 Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One.* 2012;**7**(5):e36530.
- 34 Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M. Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. *J Thorac Oncol.* 2017;**12**(6):943–53.
- 35 Ge X, Liu Z, Weng S, Xu H, Zhang Y, Liu L, et al. Integrative pharmacogenomics revealed three subtypes with different immune landscapes and specific therapeutic responses in lung adenocarcinoma. *Comput Struct Biotechnol J.* 2022;**20**:3449–60.
- 36 Roh W, Geffen Y, Cha H, Miller M, Anand S, Kim J, et al. High-resolution profiling of lung adenocarcinoma identifies expression subtypes with specific biomarkers and clinically relevant vulnerabilities. *Cancer Res.* 2022;**82**(21):3917–31.
- 37 Tamborero D, Rubio-Perez C, Muiños F, Sabarinathan R, Piulats JM, Muntasell A, et al. A Pan-cancer landscape of interactions between solid tumors and infiltrating immune cell populations. *Clin Cancer Res.* 2018;**24**(15):3717–28.
- 38 Bodor JN, Kasireddy V, Borghaei H. First-line therapies for metastatic lung adenocarcinoma without a driver mutation. *J Oncol Pract.* 2018;**14**(9):529–35.
- 39 Remon J, Majem M. EGFR mutation heterogeneity and mixed response to EGFR tyrosine kinase inhibitors

- of non small cell lung cancer: a clue to overcoming resistance. *Transl Lung Cancer Res.* 2013;**2**(6):445–8.
- 40 Garon EB, Rizvi NA, Hui R, Leigh N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non–small-cell lung cancer. *N Engl J Med.* 2015;**372**(21):2018–28.
- 41 Lagos GG, Izar B, Rizvi NA. Beyond tumor PD-L1: emerging genomic biomarkers for checkpoint inhibitor immunotherapy. *Am Soc Clin Oncol Educ Book.* 2020;**40**:1–11.
- 42 Peters S, Dziadziuszko R, Morabito A, Felip E, Gadgeel SM, Cheema P, et al. Atezolizumab versus chemotherapy in advanced or metastatic NSCLC with high blood-based tumor mutational burden: primary analysis of BFAST cohort C randomized phase 3 trial. *Nat Med.* 2022;**28**(9):1831–9.
- 43 Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med.* 2018;**24**(10):1550–8.
- 44 Qiu Y, Chen T, Hu R, Zhu R, Li C, Ruan Y, et al. Next frontier in tumor immunotherapy: macrophage-mediated immune evasion. *Biomark Res.* 2021;**9**(1):72.
- 45 Battle E, Massagué J. Transforming growth factor- β signaling in immunity and cancer. *Immunity.* 2019;**50**(4):924–40.
- 46 Tawbi HA, Schadendorf D, Lipson EJ, Ascierto PA, Matamala L, Castillo Gutiérrez E, et al. Relatlimab and nivolumab versus nivolumab in untreated advanced melanoma. *N Engl J Med.* 2022;**386**(1):24–34.
- 47 Felip E, Majem M, Doger B, Clay TD, Carcereny E, Bondarenko I, et al. A phase II study (TACTI-002) in first-line metastatic non–small cell lung carcinoma investigating efitilagimod alpha (soluble LAG-3 protein) and pembrolizumab: updated results from a PD-L1 unselected population. *J Clin Oncol.* 2022;**40**(16 Suppl):9003.
- Fig. S1.** Flow diagram of included gene expression datasets search and filtering criteria for this study.
- Fig. S2.** Computational framework for LUAD consensus subtype definition.
- Fig. S3.** Relative activity levels (GSVA scores) of the 50 studied landmark pathways across LUAD subtypes.
- Fig. S4.** Overall survival between subtypes associated with better prognosis and worse prognosis in the analysis by individual LUAD subtype.
- Fig. S5.** Correlation between pathway profiling-based subtypes and Wilkerson et al.'s mRNA-based subtypes.
- Fig. S6.** Immune cell lines relative abundance across LUAD subtypes.
- Fig. S7.** Immune checkpoints expression across LUAD subtypes.
- Fig. S8.** Relative activity levels of the fifty studied pathways in each of the 111 CPTAC-3 LUAD samples that were assigned to a consensus subtype.
- Fig. S9.** LUAD cancer cell lines (LUAD-CCL) used for the potential treatment strategies discovery analysis.
- Table S1.** List of gene expression datasets included in this study.
- Table S2.** Correlation of dataset ids with LUAD subtypes.
- Table S3.** FDR-adjusted p values for pairwise comparisons of TMB values between LUAD subtypes.
- Table S4.** Percentage of positive patients for each single nucleotide base substitution mutational signature across LUAD subtypes.
- Table S5.** FDR-adjusted p values for pairwise comparisons of copy number rate values between LUAD subtypes.
- Table S6.** FDR-adjusted p values for pairwise comparisons of copy number rate values between LUAD subtypes.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

6. DISCUSSION

6.1 Transcriptional profiling of molecular pathways yields a robust classification of lung adenocarcinoma and lung squamous cell carcinoma

Clinical management of patients with NSCLC was mainly based on tumor staging and pathological diagnosis. However, during the last few years molecular subclassification and genomic profiling of tumors has become crucial, especially for patients diagnosed with lung adenocarcinoma. In this way, NSCLC classification has evolved from a morphological standpoint to a more granular categorization by incorporating molecular features and new knowledge from translational research.

In this context, tumors are currently screened for specific genomic alterations that can predict survival benefit and sensitivity to targeted therapies. These genomic alterations are enriched in lung adenocarcinoma, but may be found in any histological subtype, and therefore, the presence of these mutations may have a greater impact on treatment decisions than the histological subtype alone. This personalized approach to treatment, also known as precision medicine, is becoming increasingly important in the management of NSCLC and other solid tumors (112).

DNA alterations do not fully capture the complexity of a tumor and its potential interactions with specific treatments. Tumors are dynamic entities composed of multiple cell types and microenvironments, and genetic mutations are only one aspect of their biology. We believe that to establish more precise and effective therapeutic approaches for NSCLC, tumors will have to be characterized in a more accurate and comprehensive way. This may include not only DNA sequencing but also analyzing the tumor's microenvironment, transcriptomics, epigenetics, and proteomics (59,67).

In the context of cancer, the expression of specific genes can influence the behavior of tumor cells, including their growth, proliferation, migration, and resistance to therapy. Understanding the patterns of gene expression in tumor cells can provide insights into the underlying mechanisms driving the development and progression of cancer and can also guide the development of new therapies. In this way, since the 2000s many different research groups have implemented whole-transcriptome gene expression profiling, coupled with bioinformatics analyses, to further classify NSCLC most represented histological subtypes (i.e., LUAD and LUSC). More interestingly, these intrinsic subtypes have also been shown to correlate with genomic features, different immune landscapes and have an impact on patients' prognosis. However, despite all the efforts, LUAD and LUSC transcriptional-based classifications have never been translated into the clinical practice. The most important

limitations of previous transcriptome-based classifications of NSCLC were their low reproducibility and the lack of overlap between the gene signatures defining each tumor subtype (69).

Our study included a broad range of more than 4,500 LUAD and 2,000 LUSC, and aimed to develop a bioinformatics framework that allows a reproducible and comprehensive classification of NSCLC. To our knowledge, no previous attempt of transcriptional-based classification has been derived from such a large set of samples.

Previous LUAD and LUSC transcriptional-based classifications were mainly based on the expression levels of individual genes, which can be affected by multiple factors, like tissue sample selection, experimental variability, and the use of different sequencing platforms (69). In our work, we quantified the activity levels of 50 landmark molecular pathways in each tumor sample using Gene Set Variation Analysis (GSVA) method (68,113). By analyzing pathways instead of individual genes, the stochastic variation of single gene expression measures is less likely to affect the results. This is because the variation in the expression of one gene within a pathway is likely to be balanced out by the expression of other genes within the same pathway, which contributes to increase the robustness of the data (68). Also, one of the main advantages of GSVA is that it can be used to analyze gene expression data from multiple datasets, and it has been shown to be more

effective at reducing batch effects than other gene expression deconvolution methods (92).

Nevertheless, GSVA scores are relative measures that depend on the number and the nature of the accompanying samples in the same dataset. In this context, we included a per dataset permutation-based procedure step across 100 iterations in our framework (**see Methods section in both published articles**). By randomly splitting each dataset in each iteration we reduce the impact of any potential sources of bias or dependence that may exist in the original data set. Notably, we were able to validate our classification using an independent set of primary tumor samples (i.e., CPTAC-LUAD and CPTAC-LUSC datasets, respectively) and NSCLC-CCLs, used for the identification of potential drug candidates (59,67,102,106). This is important as it allows to confirm that the method is not just fitting to the original set of samples, but it could be tested and validated in prospective independent datasets.

6.2 Pathway transcriptional profiling-based classification correlates and further subdivides widely accepted Wilkerson et al.'s mRNA-based subtypes

The Wilkerson et al. LUSC and LUAD mRNA-based classification model, which was then also validated and adopted in the 2012 and 2014 famous TCGA comprehensive characterization of LUSC and LUAD tumors, respectively, was a seminal work, as it improved the understanding of the

biological mechanisms underlying LUAD and LUSC intrinsic molecular heterogeneity (52,55,64,65).

We evaluated the correlation between the Wilkerson et al. model and the present classification. It is important to emphasize that our sample size far exceeds the datasets used to establish Wilkerson et al.'s mRNA subtypes, both for LUAD and LUSC, which increases tumor diversity and the chances of identifying previously unreported subtypes. Also, the methodology behind previous mRNA-based subtypes definition is differential gene expression of individual genes, which, as mentioned in 7.1, is more susceptible to stochastic sources of variation.

Overall, we found that some of the LUAD and LUSC pathway transcriptional profiling-based subtypes that we identified overlap with the intrinsic subtypes described by Wilkerson et al. However, our classification further subdivides the one proposed by Wilkerson et al. (**Figure 15**). For example, immune-enriched mRNA-based secretory LUSC subtype, better aligned with our SCC2 and SCC3 pathway transcriptional profiling-based subtypes, consistent with their also higher immune infiltration. However, SCC2 and SCC3 present different proliferation-related gene expression patterns. Moreover, our SCC1 and SCC5 LUSC and AD7 LUAD pathway transcriptional profiling-based subtypes were not found to be particularly associated with any of the Wilkerson et al. LUSC and LUAD mRNA subtypes,

respectively. This further subdivision of the Wilkerson et al. classification, also observed for other LUSC and LUAD mRNA-based subtypes, and the identification of potential new groups, suggests an increased resolution of our classification.

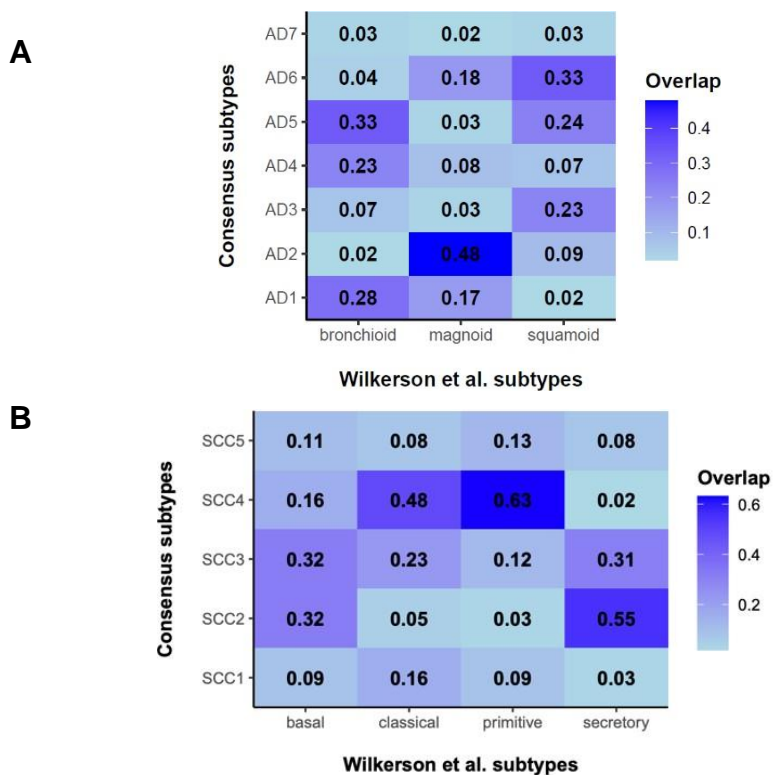


Figure 15. Links between pathway transcriptional profiling-based subtypes and Wilkerson et al.'s mRNA-based subtypes.

(A) Wilkerson et al.'s LUAD mRNA- based subtypes (i.e., bronchioid, magnoid and squamoid) were assigned to each of our LUAD samples using the nearest centroid predictor approach described by the original authors. The overlap degree (proportion of samples within the same category) between the two classifications is displayed. **(B)** Wilkerson et al.'s LUSC mRNA-based subtypes (i.e., secretory, basal, classical, and primitive) were assigned to each of our LUSC samples using the nearest centroid predictor. The overlap degree (proportion of samples within the same category) between the two classifications is displayed.

6.3 Integration of transcriptomic and genomic data could improve current NSCLC patient stratification in patients with driver positive lung adenocarcinoma

Genomic profiling is currently crucial for identifying actionable driver alterations and guiding treatment in patients with NSCLC, especially for LUAD, in which the frequency of driver actionable alterations far exceeds that of LUSC (4). Moreover, recent research has shown that the presence of multiple concurrent genetic mutations could also play a role in determining treatment response (16). However, LUAD is a highly heterogeneous disease, and the presence of certain co-alterations alone does not fully explain the varied treatment responses seen in patients with the same driver mutation.

Our research showed that while there are associations between certain genetic mutations and specific pathway transcriptional-based LUAD subtypes (i.e., *EGFR*, *TP53*, *STK11*), patients with the same mutation can be found across different transcriptional subtypes. *KRAS* mutant LUAD was found to be particularly heterogeneous in this regard. Thus, in addition to concurrent genomic alterations, differences in the activation of transcriptional pathways could explain that patients harboring specific driver alterations would have distinct clinical outcomes upon targeted therapy.

Conversely, we observed that LUAD with different driver alterations were allocated within the same transcriptional

subtype. This could suggest that different mutations may give rise to similar transcriptional phenotypes, which could potentially be treated with similar combinatorial therapies.

6.4 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes show differential genome instability features

Genomic instability, which is a common feature of cancer cells, is driven by both DNA damage and errors made by the DNA damage repair (DDR) systems (114). On the one hand, cigarette smoking contributes significantly to the accumulation of DNA damage and is the most important risk factor for lung cancer development (115). Indeed, tobacco-related genomic signature (SBS4) was found to be overrepresented across both LUAD and LUSC pathway transcriptional profiling-based subtypes. However, no relevant differences were observed between LUAD or LUSC subtypes in terms of the presence of SBS4 signature, which suggests an equivalent tobacco carcinogens-related DNA damage in LUAD and LUSC development.

On the other hand, we did discover differences in terms of tumor mutational burden and copy number alterations between subtypes. SCC1 and SCC4 LUSC subtypes and AD2 and AD6 LUAD subtypes were enriched for copy number alterations (CNA) rates compared with other subtypes, respectively.

Moreover, AD2 and AD6 tumors further displayed higher TMB values when compared with other LUAD subtypes.

Interestingly, SCC1, SCC4, AD2 and AD6 also demonstrated higher activation of proliferation-related pathways and a potentially higher impairment of DDR mechanisms.

These results suggest that the greater genomic instability found for some of the subtypes (i.e., LUAD: AD2, AD6; LUSC: SCC1, SCC4), might not be the result of a higher exposure to exogenous carcinogens (i.e., tobacco), but rather a consequence of higher error rates made by DDR mechanisms, or replication stress due to high proliferation rates (116). Moreover, AD2 and AD6 LUAD subtypes also displayed higher frequency of genomic *TP53* alterations, which have also been associated with higher genome instability (117).

6.5 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes display specific therapeutic vulnerabilities

As previously mentioned, the lack of association of previously described NSCLC transcriptional subtypes with specific therapeutic vulnerabilities has prevented their use in the clinic (69). Here, we tried to overcome this limitation by integrating drug sensitivity data from *in vitro* drug sensitivity studies (i.e., CTRPv2, GDSC, PRISM) (118). These large projects have greater drug coverage than other available databases such as CMap, which has been previously used for similar purposes (60). Although we are aware of the limitations of cancer cell lines drug screening studies and that there can be discrepancies with clinical benefit in patients, we identified

some potential therapeutic vulnerabilities for specific LUAD and LUSC subtypes.

In this way, chemotherapy alone, or combined with immunotherapy, is still widely used for the treatment of driver negative NSCLC. However, clinical responses upon chemotherapy regimens are highly heterogeneous. Therefore, identifying which patients may derive larger benefit or experience an early progression from these treatments remains an important clinical need. Regarding LUAD subtypes, results showed that AD2, AD3, AD6 and AD7 cell lines might benefit from G2/M or S/G2 chemotherapies (i.e., vincristine, gemcitabine). This is in line with the higher expression of proliferation related pathways in these groups. Moreover, cancer cells classified within AD2 showed potential sensitivity to AZD7762 CHEK1 inhibitor, which correlates with the higher genome instability seen in this subtype. It is important to note that this was conceived as a drug repositioning exercise to identify potential therapeutic specificities for the subtypes. Thus, some of the drugs may not be currently used as a treatment in lung adenocarcinoma (i.e., vincristine). However, drugs from the same family, such as vinorelbine could be considered.

Similarly, LUSC cancer cell lines classified in SCC4 genome unstable (i.e., higher CNA rate) and proliferative subtype showed potential sensitivity to different G2M phase chemotherapy regimens (i.e., docetaxel, paclitaxel, and

vincristine). In this context, replication stress might cause these tumors to be dependent on the DDR machinery (cell cycle checkpoints and DNA repair mechanisms) for survival, which is in concordance with the observed overexpression of DNA repair pathways and G2M checkpoints observed for this LUSC subtype (116). Furthermore, SCC1 and SCC4 cell lines also displayed potential sensitivity to some DNA damage and cell cycle targeted therapies, which is also in line with the proliferative nature that characterize these subtypes.

6.6 Lung adenocarcinoma and lung squamous cell carcinoma transcriptional-based subtypes display different immune landscapes with potential therapeutic implications

Immunotherapy alone or in combination with chemotherapy is the standard of care for metastatic NSCLC without actionable driver alterations (25). However, only a minority of patients (~20%) are long-term survivors and experience long-lasting responses to immunotherapy (119).

The only biomarker utilized in clinical practice is tumor PD-L1 expression. Other potential markers like TMB or certain genomic alterations (e.g., *PTEN*, *STK11*, *KEAP1*, and *TP53*) have also been considered potential biomarkers for immunotherapy, but they are not able to accurately predict clinical benefit. The immune system is indeed a complex and

intricate network of intercellular interactions, and it is challenging to pinpoint a single factor that determines its activity. This complexity makes it difficult to achieve a situation similar to molecularly targeted therapy, where treatment efficacy is based on a single driving alteration, making patient selection for immunotherapy regimens extraordinarily difficult (120).

In this context, recent studies have demonstrated the capacity of gene expression profiling to provide comprehensive and clinically relevant information beyond single biomarkers (92–94). Given the intrinsic biological differences observed for the different pathway transcriptional profiling-based LUAD and LUSC subtypes, we were interested in elucidating whether subtype-specific transcriptional footprints might shape different immune landscapes. Understanding these differences may contribute to the selection of patients that might benefit from immunotherapy treatment.

We used multiple gene expression signatures, each one representing a specific immune cell type, to infer the immune cell infiltration pattern of the different subtypes. Results from these analyses revealed different immune cell infiltration patterns for LUAD and LUSC transcriptional subtypes. We observed that the presence of anti-tumoral immune response (i.e., cytotoxic cells, Th1 cells, B cells) can coexist with immunosuppressive cells (i.e., T regulatory cells, macrophages M2), which act as a major barrier to cancer

immunotherapy (121). For instance, AD3, enriched in tumors with relatively high cytotoxic T cell infiltration and PD-L1 gene expression, was not specially associated with ICI therapy response (only 12.5% of AD3 tumors were predicted as responders). Importantly, AD3 tumors co-expressed other immune checkpoint inhibitors and T-cell exhaustion markers. Moreover, AD3 tumors were also enriched in tumors with high infiltration M2 macrophages and T regulatory cells, which could contribute to immune evasion and intrinsic resistance to immunotherapy in AD3 tumors. The same infiltration patterns, combining both anti-tumoral and pro-tumoral immune cells expression, were observed for LUSC SCC2 and SCC3 subtypes. For this reason, rational combinations of ICI and immune cell specific targeted therapies could probably improve clinical outcomes in solid tumors.

Although further validation through other techniques that provide greater cellular resolution (i.e., scRNA-seq) would be useful, these results support the idea that a comprehensive characterization of immune contexture could improve our capability to predict response or resistance to immunotherapy, as opposed to testing single biomarkers (i.e., PD-L1, TMB). Moreover, better patients' stratification might help clinicians to define more rational combinations of distinct ICIs and other treatments targeting immunosuppressive immune cells. Unfortunately, most clinical trials are focusing on single biomarkers such as PD-L1 and just combining distinct

immunotherapies without conducting any patient selection or evaluating tumor's immune contexture.

6.7 Limitations of the present work and future perspectives

Despite the robustness of our classification, this study has some limitations. First, this is a retrospective analysis of a wide collection of **microarrays** and **RNA-Seq** gene expression datasets, which rely on fresh tissue biopsies to ensure mRNA quality. Thus, further efforts are needed towards the implementation of this classification framework in formalin-fixed paraffin-embedded tumor samples, which are more likely to be available in the clinical setting than fresh tissue. For instance, we believe that with the incorporation of new profiling technologies, that allow whole-transcriptome gene expression profiling in FFPE samples, it will be possible to evaluate the clinical relevance of our framework using clinical samples. Most datasets used included early-stage patients who were surgically resected and did not receive systemic therapy, or this information was not available in the associated clinical information. For instance, this is particularly relevant for the results regarding immunotherapy response predictions, which should be further validated in retrospective and prospective studies of patients with LUAD and LUSC treated with ICI. In addition, all the results regarding the association with genomic features (i.e., TMB, mutational signatures, CNA, and DDR deficiency) were based on the TCGA-LUAD and TCGA-LUSC

datasets, respectively, since the rest of studies lacked associated whole exome sequencing data. However, the CPTAC-LUAD and CPTAC-LUSC validation datasets allowed us to prove that our classification can be reproduced in an independent dataset and that the association of the different LUAD and LUSC subtypes with copy number alterations and TMB is also conserved beyond the gene expression patterns used to classify these samples. Nevertheless, validation in a prospective cohort is warranted to further validate the clinical utility of this classification.

Also related with the association between the subtypes and the presence of relevant oncogenic drivers, especially in the case of LUAD, it would have also been interesting to stratify patients by specific gene mutation (i.e., *KRAS G12C*, *KRAS G12D*...etc.). Unfortunately, this information was not available for most of the gene expression datasets used in this study (only in TCGA-LUAD or TCGA-LUSC), preventing this analysis.

Regarding cancer cell lines drug sensitivity results, potential drug candidates are based on *in vitro* data and, therefore, do not take into consideration the interplay between cancer cells and TME. However, these preclinical models are continuously used in experimental research for similar purposes (i.e., drug screening and hypothesis generation) and we believe that this exercise could be useful to prioritize which compounds could be tested in more advanced preclinical models (i.e., tumoroids and patient-

derived xenografts). Thus, further validation in more advanced *in vivo* models and prospective patient cohorts is warranted to confirm these results but is beyond the scope of this dissertation.

Finally, have recently been important developments in the field of transcriptomics to increase resolution and specificity. In traditional RNA-seq experiments, gene expression is measured from a homogenized mixture of cells, resulting in the loss of spatial information. In contrast, single-cell RNA-Seq allow the determination of the gene expression profiles of individual tumor cells and microenvironment populations. Also, spatial transcriptomics enables the mapping of gene expression profiles onto the tissue architecture, providing a comprehensive understanding of gene expression patterns in their native spatial context. In this way, these two high-throughput techniques provide information about tumor heterogeneity and cellular interactions. This would have been particularly useful for the study of the immune landscape and TME heterogeneity, which in the end shapes immunotherapy responses. Future studies in this direction are beginning to revolutionize our understanding of tissue biology and to drive the development of new therapies and diagnostic tools in a more rational way.

7. CONCLUSIONS

In this work, we have presented and validated a robust, reproducible, and clinically relevant classification of LUAD and LUSC, based on the activity levels of landmark molecular pathways. To our knowledge, no previous LUAD or LUSC classification has been derived from a such a large collection of tumor samples. Despite significant difficulties, we believe that the joint implementation of transcriptomic and genomic data could improve patient stratification and may pave the way for guiding personalized-medicine approaches in patients with NSCLC.

Specifically, we have shown that:

- 1) Transcriptional profiling of landmark molecular pathways in a large tumor sample collection yielded seven and five pathway transcriptional profiling-based subtypes in LUAD and LUSC, respectively;
- 2) Cox multivariate analyses adjusted for sex, age, stage, smoking history, and study clinical covariates, reveal a significant association between LUAD subtypes and overall survival;
- 3) Analysis of available genomic alterations revealed an association between LUAD transcriptional subtypes and specific actionable oncogenic alterations;
- 4) LUAD and LUSC transcriptional subtypes showed different genomic features, specifically in terms of tumor mutational

burden, copy number alteration rates and DNA damage repair capacity;

- 5) Integration of *in vitro* drug sensitivity data from large pharmacogenomic studies, allowed the identification of specific therapeutic vulnerabilities for these subtypes, in line with their molecular characteristics;
- 6) Gene expression profiling of immune-related gene expression signatures and biomarkers revealed different immune landscapes for the LUAD and LUSC subtypes that might help improving our capability to predict response to immunotherapy, as opposed to testing single biomarkers (i.e., PD-L1, TMB), which have shown limited capacity to accurately predict clinical benefit;
- 7) Integration of genomic and transcriptomic data might improve patient stratification and treatment guidance towards more personalized-medicine approaches.

Bibliography

1. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res.* august 2014;3(4):242-9.
2. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science.* 4 june 2004;304(5676):1497-500.
3. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *N Engl J Med.* 20 may 2004;350(21):2129-39.
4. Hendriks LE, Kerr KM, Menis J, Mok TS, Nestle U, Passaro A, et al. Oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up☆. *Ann Oncol* [Internet]. 24 january2023 [citat 7 march 2023];0(0). Disponible a: [https://www.annalsofncology.org/article/S0923-7534\(22\)04781-0/fulltext](https://www.annalsofncology.org/article/S0923-7534(22)04781-0/fulltext)
5. Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol Off J Eur Soc Med Oncol.* 1 october 2018;29(Suppl 4):iv192-237.
6. Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N Engl J Med.* 13 august 2020;383(7):640-9.
7. Tan AC, Tan DSW. Targeted Therapies for Lung Cancer Patients With Oncogenic Driver Molecular Alterations. *J Clin Oncol.* 20 february 2022;40(6):611-25.
8. Meador CB, Hata AN. Acquired resistance to targeted therapies in NSCLC: Updates and evolving insights. *Pharmacol Ther.* june 2020;210:107522.
9. Shields MD, Marin-Acevedo JA, Pellini B. Immunotherapy for Advanced Non–Small Cell Lung Cancer: A Decade of Progress. *Am Soc Clin Oncol Educ Book.* june 2021;(41):e105-27.
10. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med.* 21 may 2015;372(21):2018-28.
11. Lagos GG, Izar B, Rizvi NA. Beyond Tumor PD-L1: Emerging Genomic Biomarkers for Checkpoint Inhibitor Immunotherapy. *Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meet.* march 2020;40:1-11.

12. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 4 february 2021;
13. Cancer of the Lung and Bronchus - Cancer Stat Facts [Internet]. SEER. [cited 21 october 2022]. Available at: <https://seer.cancer.gov/statfacts/html/lungb.html>
14. Pirker R. Conquering lung cancer: current status and prospects for the future. *Pulmonology.* 1 september 2020;26(5):283-90.
15. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer.* september 2015;10(9):1243-60.
16. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med.* march 2011;17(3):297-303.
17. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer.* february 2011;6(2):244-85.
18. Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, et al. Screening for Epidermal Growth Factor Receptor Mutations in Lung Cancer. *N Engl J Med.* 3 september 2009;361(10):958-67.
19. Li T, Kung HJ, Mack PC, Gandara DR. Genotyping and Genomic Profiling of Non-Small-Cell Lung Cancer: Implications for Current and Future Therapies. *J Clin Oncol.* 10 march 2013;31(8):1039-49.
20. Finn SP, Addeo A, Dafni U, Thunnissen E, Bubendorf L, Madsen LB, et al. Prognostic Impact of KRAS G12C Mutation in Patients With NSCLC: Results From the European Thoracic Oncology Platform Lungscape Project. *J Thorac Oncol.* 1 june 2021;16(6):990-1002.
21. Clinical Practice Living Guidelines – Metastatic Non-Small-Cell Lung Cancer | ESMO [Internet]. [citat 21 october 2022]. Disponible a: <https://www.esmo.org/guidelines/guidelines-by-topic/lung-and-chest-tumours/clinical-practice-living-guidelines-metastatic-non-small-cell-lung-cancer>
22. Paik PK, Pillai RN, Lathan CS, Velasco SA, Papadimitrakopoulou V. New Treatment Options in Advanced Squamous Cell Lung Cancer. *Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meet.* january2019;39:e198-206.

23. Tsao MS, Nicholson AG, Maleszewski JJ, Marx A, Travis WD. Introduction to 2021 WHO Classification of Thoracic Tumors. *J Thorac Oncol*. 1 january2022;17(1):e1-4.
24. Lam VK, Tran HT, Banks KC, Lanman RB, Rinsurongkawong W, Peled N, et al. Targeted Tissue and Cell-Free Tumor DNA Sequencing of Advanced Lung Squamous-Cell Carcinoma Reveals Clinically Significant Prevalence of Actionable Alterations. *Clin Lung Cancer*. 1 january2019;20(1):30-36.e3.
25. Ettinger DS, Aisner DL, Wood DE, Akerley W, Bauman J, Chang JY, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 5.2018. *J Natl Compr Cancer Netw JNCCN*. july 2018;16(7):807-21.
26. Ribas A. Adaptive immune resistance: How cancer protects from immune attack. *Cancer Discov*. september 2015;5(9):915-9.
27. Pawelczyk K, Piotrowska A, Ciesielska U, Jablonska K, Glatzel-Plucinska N, Grzegorzolka J, et al. Role of PD-L1 Expression in Non-Small Cell Lung Cancer and Their Prognostic Significance according to Clinicopathological Factors and Diagnostic Markers. *Int J Mol Sci*. 14 february 2019;20(4):824.
28. Grant MJ, Herbst RS, Goldberg SB. Selecting the optimal immunotherapy regimen in driver-negative metastatic NSCLC. *Nat Rev Clin Oncol*. october 2021;18(10):625-44.
29. Kazandjian D, Suzman DL, Blumenthal G, Mushti S, He K, Libeg M, et al. FDA Approval Summary: Nivolumab for the Treatment of Metastatic Non-Small Cell Lung Cancer With Progression On or After Platinum-Based Chemotherapy. *The Oncologist*. may 2016;21(5):634-42.
30. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med*. 9 july 2015;373(2):123-35.
31. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N Engl J Med*. 22 october 2015;373(17):1627-39.
32. Borghaei H, Gettinger S, Vokes EE, Chow LQM, Burgio MA, de Castro Carpeno J, et al. Five-Year Outcomes From the Randomized, Phase III Trials CheckMate 017 and 057: Nivolumab Versus Docetaxel in Previously Treated Non-Small-Cell Lung Cancer. *J Clin Oncol Off J Am Soc Clin Oncol*. 1 march 2021;39(7):723-33.
33. Lim ZF, Ma PC. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *J Hematol Oncol J Hematol Oncol*. 9 december 2019;12(1):134.

34. Rotow J, Bivona TG. Understanding and targeting resistance mechanisms in NSCLC. *Nat Rev Cancer*. november 2017;17(11):637-58.
35. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. november 2015;21(11):1350-6.
36. Horr C, Buechler SA. Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression. *Npj Breast Cancer*. 12 october 2021;7(1):1-13.
37. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*. 20 november 2001;98(24):13784-9.
38. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*. 20 november 2001;98(24):13790-5.
39. Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*. august 2002;8(8):816-24.
40. Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, et al. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene*. july 2004;23(31):5360-70.
41. Inamura K, Fujiwara T, Hoshida Y, Isagawa T, Jones MH, Virtanen C, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*. october 2005;24(47):7105-13.
42. Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, et al. Gene Expression Profiling Reveals Reproducible Human Lung Adenocarcinoma Subtypes in Multiple Independent Patient Cohorts. *J Clin Oncol*. november 2006;24(31):5079-90.
43. Takeuchi T, Tomida S, Yatabe Y, Kosaka T, Osada H, Yanagisawa K, et al. Expression Profile–Defined Classification of Lung Adenocarcinoma Shows Close Relationship With Underlying Major Genetic Changes and Clinicopathologic Behaviors. *J Clin Oncol*. 20 april 2006;24(11):1679-88.
44. Shibata T, Hanada S, Kokubu A, Matsuno Y, Asamura H, Ohta T, et al. Gene expression profiling of epidermal growth factor receptor/KRAS pathway activation in lung adenocarcinoma. *Cancer Sci*. 2007;98(7):985-91.

45. Park YY, Park ES, Kim SB, Kim SC, Sohn BH, Chu IS, et al. Development and Validation of a Prognostic Gene-Expression Signature for Lung Adenocarcinoma. *PLOS ONE*. 7 september 2012;7(9):e44225.
46. Staaf J, Jönsson G, Jönsson M, Karlsson A, Isaksson S, Salomonsson A, et al. Relation between smoking history and gene expression profiles in lung adenocarcinomas. *BMC Med Genomics*. 7 june 2012;5(1):22.
47. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential Pathogenesis of Lung Adenocarcinoma Subtypes Involving Sequence Mutations, Copy Number, Chromosomal Instability, and Methylation. *PLOS ONE*. 10 may 2012;7(5):e36530.
48. Cheung WKC, Zhao M, Liu Z, Stevens LE, Cao PD, Fang JE, et al. Control of Alveolar Differentiation by the Lineage Transcription Factors GATA6 and HOPX Inhibits Lung Adenocarcinoma Metastasis. *Cancer Cell*. 10 june 2013;23(6):725-38.
49. Fukui T, Shaykhiev R, Agosto-Perez F, Mezey JG, Downey RJ, Travis WD, et al. Lung adenocarcinoma subtypes based on expression of human airway basal cell genes. *Eur Respir J*. 1 november 2013;42(5):1332-44.
50. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. july 2014;511(7511):543-50.
51. Ringnér M, Staaf J. Consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. *Oncotarget*. 16 july 2016;7(33):52957-73.
52. Chen F, Zhang Y, Parra E, Rodriguez J, Behrens C, Akbani R, et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene*. march 2017;36(10):1384-93.
53. Hu F, Zhou Y, Wang Q, Yang Z, Shi Y, Chi Q. Gene Expression Classification of Lung Adenocarcinoma into Molecular Subtypes. *IEEE/ACM Trans Comput Biol Bioinform*. july 2020;17(4):1187-97.
54. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell*. 9 july 2020;182(1):200-225.e35.
55. Ge X, Liu Z, Weng S, Xu H, Zhang Y, Liu L, et al. Integrative pharmacogenomics revealed three subtypes with different immune landscapes and specific therapeutic responses in lung adenocarcinoma. *Comput Struct Biotechnol J*. 2022;20:3449-60.

56. Connectivity Map (CMAP) [Internet]. Broad Institute. 2015 [citat 21 october 2022]. Disponible a: <https://www.broadinstitute.org/connectivity-map-cmap>
57. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, et al. Gene Expression Signatures for Predicting Prognosis of Squamous Cell and Adenocarcinomas of the Lung. *Cancer Res.* 2 august 2006;66(15):7466-72.
58. Larsen JE, Pavey SJ, Passmore LH, Bowman R, Clarke BE, Hayward NK, et al. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis.* 1 march 2007;28(3):760-6.
59. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types. *Clin Cancer Res.* 29 september 2010;16(19):4864-75.
60. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature.* september 2012;489(7417):519-25.
61. Brambilla C, Laffaire J, Lantuejoul S, Moro-Sibilot D, Mignotte H, Arbib F, et al. Lung Squamous Cell Carcinomas with Basaloid Histology Represent a Specific Molecular Entity. *Clin Cancer Res.* 13 november 2014;20(22):5777-86.
62. Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell.* 5 august 2021;184(16):4348-4371.e40.
63. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 23 december 2015;1(6):417-25.
64. Hijazo-Pechero S, Alay A, Marín R, Vilariño N, Muñoz-Pinedo C, Villanueva A, et al. Gene Expression Profiling as a Potential Tool for Precision Oncology in Non-Small Cell Lung Cancer. *Cancers.* january2021;13(19):4734.
65. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn.* 1 may 2015;17(3):251-64.
66. Fernandes MGO, Jacob M, Martins N, Moura CS, Guimarães S, Reis JP, et al. Targeted Gene Next-Generation Sequencing Panel in

Patients with Advanced Lung Adenocarcinoma: Paving the Way for Clinical Implementation. *Cancers*. september 2019;11(9):1229.

67. Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. *Nature*. august 2017;548(7667):297-303.

68. Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat Med*. may 2019;25(5):751-8.

69. Tuxen IV, Rohrberg KS, Oestrup O, Ahlborn LB, Schmidt AY, Spanggaard I, et al. Copenhagen Prospective Personalized Oncology (CoPPO)—Clinical Utility of Using Molecular Profiling to Select Patients to Phase I Trials. *Clin Cancer Res*. 15 february 2019;25(4):1239-47.

70. Pleasance E, Bohm A, Williamson LM, Nelson JMT, Shen Y, Bonakdar M, et al. Whole-genome and transcriptome analysis enhances precision cancer treatment options. *Ann Oncol*. 1 september 2022;33(9):939-49.

71. Neel DS, Bivona TG. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. *Npj Precis Oncol*. 20 march 2017;1(1):1-6.

72. Alam N, Gustafson KS, Ladanyi M, Zakowski MF, Kapoor A, Truskinovsky AM, et al. Small-cell carcinoma with an epidermal growth factor receptor mutation in a never-smoker with gefitinib-responsive adenocarcinoma of the lung. *Clin Lung Cancer*. 1 september 2010;11(5):E1-4.

73. Yamada T, Takeuchi S, Nakade J, Kita K, Nakagawa T, Nanjo S, et al. Paracrine receptor activation by microenvironment triggers bypass survival signals and ALK inhibitor resistance in EML4-ALK lung cancer cells. *Clin Cancer Res Off J Am Assoc Cancer Res*. 1 july 2012;18(13):3592-602.

74. Tan CS, Gilligan D, Pacey S. Treatment approaches for EGFR-inhibitor-resistant patients with non-small-cell lung cancer. *Lancet Oncol*. 1 september 2015;16(9):e447-59.

75. Zhu X, Chen L, Liu L, Niu X. EMT-Mediated Acquired EGFR-TKI Resistance in NSCLC: Mechanisms and Strategies. *Front Oncol*. 2019;9:1044.

76. Fumarola C, Bonelli MA, Petronini PG, Alfieri RR. Targeting PI3K/AKT/mTOR pathway in non small cell lung cancer. *Biochem Pharmacol*. 1 august 2014;90(3):197-207.

77. Coldren CD, Helfrich BA, Witta SE, Sugita M, Lapadat R, Zeng C, et al. Baseline Gene Expression Predicts Sensitivity to Gefitinib in Non-Small Cell Lung Cancer Cell Lines. *Mol Cancer Res.* 1 august 2006;4(8):521-8.
78. Balko JM, Potti A, Saunders C, Stromberg A, Haura EB, Black EP. Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics.* 10 november 2006;7(1):289.
79. Zhang Z, Lee JC, Lin L, Olivas V, Au V, LaFramboise T, et al. Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nat Genet.* august 2012;44(8):852-60.
80. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, et al. An Epithelial-Mesenchymal Transition Gene Signature Predicts Resistance to EGFR and PI3K Inhibitors and Identifies Axl as a Therapeutic Target for Overcoming EGFR Inhibitor Resistance. *Clin Cancer Res.* 2 january 2013;19(1):279-90.
81. Terai H, Soejima K, Yasuda H, Nakayama S, Hamamoto J, Arai D, et al. Activation of the FGF2-FGFR1 Autocrine Pathway: A Novel Mechanism of Acquired Resistance to Gefitinib in NSCLC. *Mol Cancer Res.* 1 july 2013;11(7):759-67.
82. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* 3 march 2014;15(3):R47.
83. Liu YN, Chang TH, Tsai MF, Wu SG, Tsai TH, Chen HY, et al. IL-8 confers resistance to EGFR inhibitors by inducing stem cell properties in lung cancer. *Oncotarget.* 18 march 2015;6(12):10415-31.
84. Rothenberg SM, Concannon K, Cullen S, Boulay G, Turke AB, Faber AC, et al. Inhibition of mutant EGFR in lung cancer cells triggers SOX2-FOXO6-dependent survival pathways. Davis R, editor. *eLife.* 16 february 2015;4:e06132.
85. Mojtabavi Naeini M, Tavassoli M, Ghaedi K. Systematic bioinformatic approaches reveal novel gene expression signatures associated with acquired resistance to EGFR targeted therapy in lung cancer. *Gene.* 15 august 2018;667:62-9.
86. Cheng C, Zhao Y, Schaafsma E, Weng YL, Amos C. An EGFR signature predicts cell line and patient sensitivity to multiple tyrosine kinase inhibitors. *Int J Cancer.* 2020;147(9):2621-33.
87. Fan XX, Wu Q. Decoding Lung Cancer at Single-Cell Level. *Front Immunol [Internet].* 2022 [citat 21 october 2022];13. Disponible a: <https://www.frontiersin.org/articles/10.3389/fimmu.2022.883758>

88. González-Silva L, Quevedo L, Varela I. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. *Trends Cancer*. 1 january2020;6(1):13-9.
89. Tamborero D, Rubio-Perez C, Muiños F, Sabarinathan R, Piulats JM, Muntasell A, et al. A Pan-cancer Landscape of Interactions between Solid Tumors and Infiltrating Immune Cell Populations. *Clin Cancer Res*. 31 july 2018;24(15):3717-28.
90. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest*. 127(8):2930-40.
91. Hwang S, Kwon AY, Jeong JY, Kim S, Kang H, Park J, et al. Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. *Sci Rep*. 20 january2020;10(1):643.
92. Rajurkar S, Mambetsariev I, Pharaon R, Leach B, Tan T, Kulkarni P, et al. Non-Small Cell Lung Cancer from Genomics to Therapeutics: A Framework for Community Practice Integration to Arrive at Personalized Therapy Strategies. *J Clin Med*. 15 june 2020;9(6):E1870.
93. Adib E, Nassar AH, Abou Alaiwi S, Groha S, Akl EW, Sholl LM, et al. Variation in targetable genomic alterations in non-small cell lung cancer by genetic ancestry, sex, smoking history, and histology. *Genome Med*. 15 april 2022;14(1):39.
94. Pakkala S, Ramalingam SS. Personalized therapy for lung cancer: striking a moving target. *JCI Insight* [Internet]. 9 august 2018 [cited 21 october 2022];3(15). Available at: <https://insight.jci.org/articles/view/120858>
95. Guo L, Chen Z, Xu C, Zhang X, Yan H, Su J, et al. Intratumoral heterogeneity of EGFR-activating mutations in advanced NSCLC patients at the single-cell level. *BMC Cancer*. 23 april 2019;19(1):369.
96. Angulo B, Suarez-Gauthier A, Lopez-Rios F, Medina P, Conde E, Tang M, et al. Expression signatures in lung cancer reveal a profile for EGFR-mutant tumours and identify selective PIK3CA overexpression by gene amplification. *J Pathol*. 2008;214(3):347-56.
97. Planck M, Isaksson S, Veerla S, Staaf J. Identification of Transcriptional Subgroups in EGFR-Mutated and EGFR/KRAS Wild-Type Lung Adenocarcinoma Reveals Gene Signatures Associated with Patient Outcome. *Clin Cancer Res*. 16 september 2013;19(18):5116-26.
98. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of Genes Upregulated in ALK-Positive and

EGFR/KRAS/ALK-Negative Lung Adenocarcinomas. *Cancer Res.* 2012;72(1):100-11.

99. Chen EY, Raghunathan V, Prasad V. An Overview of Cancer Drugs Approved by the US Food and Drug Administration Based on the Surrogate End Point of Response Rate. *JAMA Intern Med.* 1 July 2019;179(7):915-21.

100. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* 28 March 2012;483(7391):603-7.

101. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* 28 March 2012;483(7391):570-5.

102. Feng F, Shen B, Mou X, Li Y, Li H. Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. *J Genet Genomics.* 20 July 2021;48(7):540-51.

103. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer.* October 2006;6(10):813-23.

104. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* January 2013;41(Database issue):D955-961.

105. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* November 2015;5(11):1210-23.

106. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer.* February 2020;1(2):235-48.

107. DepMap: The Cancer Dependency Map Project at Broad Institute [Internet]. [cited 21 October 2022]. Available at: <https://depmap.org/portal/>

108. Trastulla L, Noorbakhsh J, Vazquez F, McFarland J, Iorio F. Computational estimation of quality and clinical relevance of cancer cell lines. *Mol Syst Biol.* July 2022;18(7):e111017.

109. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. *Cell.* 27 July 2017;170(3):564-576.e16.

110. Michelotti A, de Scordilli M, Bertoli E, De Carlo E, Del Conte A, Bearz A. NSCLC as the Paradigm of Precision Medicine at Its Finest: The Rise of New Druggable Molecular Targets for Advanced Disease. *Int J Mol Sci.* January 2022;23(12):6748.

111. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. 16 january2013;14(1):7.
112. Skoulidis F, Heymach JV. Co-occurring genomic alterations in non-small cell lung cancer biology and therapy. *Nat Rev Cancer*. september 2019;19(9):495-509.
113. Torgovnick A, Schumacher B. DNA repair mechanisms in cancer development and therapy. *Front Genet*. 23 april 2015;6:157.
114. Huang CC, Lai CY, Tsai CH, Wang JY, Wong RH. Combined effects of cigarette smoking, DNA methyltransferase 3B genetic polymorphism, and DNA damage on lung cancer. *BMC Cancer*. 29 september 2021;21(1):1066.
115. Zhu H, Swami U, Preet R, Zhang J. Harnessing DNA Replication Stress for Novel Cancer Therapy. *Genes*. september 2020;11(9):990.
116. Eischen CM. Genome Stability Requires p53. *Cold Spring Harb Perspect Med*. june 2016;6(6):a026096.
117. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacGx: an R package for analysis of large pharmacogenomic datasets. *Bioinforma Oxf Engl*. 15 april 2016;32(8):1244-6.
118. Punekar SR, Shum E, Grello CM, Lau SC, Velcheti V. Immunotherapy in non-small cell lung cancer: Past, present, and future directions. *Front Oncol [Internet]*. 2022 [cited 7 march 2023];12. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9382405/>
119. Murciano-Goroff YR, Warner AB, Wolchok JD. The future of cancer immunotherapy: microenvironment-targeting combinations. *Cell Res*. june 2020;30(6):507-19.
120. Qiu Y, Chen T, Hu R, Zhu R, Li C, Ruan Y, et al. Next frontier in tumor immunotherapy: macrophage-mediated immune evasion. *Biomark Res*. 9 october 2021;9(1):72.
121. van Maldegem F, Downward J. Mutant KRAS at the Heart of Tumor Immune Evasion. *Immunity*. 14 january2020;52(1):14-6.

