PEDECIBA BIOINFORMATIC

MASTER THESIS

A machine learning-based classification for Small Cell Lung Cancer subtypes: implications for prognosis and therapy selection

Author
Nicole Kiedanski

Supervisors
PhD. Lucia Spangenberg
PhD. Eike Staub

November 1st, 2024

Contents

1.Background **4**
2. General Goal **4**
3. Specific Goals **4**
4. Introduction **5**
4.1 Types of pulmonary carcinomas 5
4.2 Neuroendocrine cells 5
4.3 Lung Neuroendocrine Neoplasms 5
4.4. Small Cell Lung Cancer 7
4.4.1 Disease characteristics 7
4.4.2 Treatment strategies 7
4.4.3 Tumor morphological and molecular characteristics 7
4.4.4 SCLC subtypes 8
4.4.5 Cell of origin 10
4.4.6 Intratumor heterogeneity 10
4.4.7 Method for subtype classification 10
4.4.8 Therapeutic Implications 11
4.5. Gene expression overview 12
4.5.1 The foundations of gene expression 12
4.5.2 Next Generation Sequencing platforms 12
4.5.3 RNA-Seq Data Analysis Workflow 13
4.5.4 Gene expression signatures 13
4.6 Machine Learning Overview 13
4.6.1 Supervised Learning 13
4.6.2 Classification problems 13
4.6.3 Nearest Centroid 14
4.6.4 K-Nearest Neighbors 14
4.6.5 Support Vector Machine 14
4.6.6 Random Forest 16
4.6.7 Performance Metrics 17
4.6.8 Cross Validation 19
5. Datasets and methods **20**
5.1. Datasets 20
5.1.1 Tempus Dataset 20
5.1.2 Cancer Cell Line Encyclopedia (CCLE) Dataset 21
5.2. Subtype Assignment based on Transcriptomic data 21

5.2.1 Tempus Labeling 21
5.2.2 CCLE Labeling 21
5.3. Machine Learning Classifier Development 21
5.3.1 Feature Selection 23
5.3.2 Hyperparameter tuning 23
5.3.3 Model Selection and Further Evaluation 24
5.4. Additional expression patterns by subtype 26
5.5. Therapeutic implications by subtype 26
6. Results and discussion 28
6.1 Subtype Assignment based on master TFs expression 28
6.2 Descriptive analysis of clinical features in Tempus SCLC cohort 28
6.3 Machine Learning Results 29
6.3.1 Grid of downstream programs used as predictors 29
6.3.2 Comparison of algorithms performances 29
6.3.4 Benchmarking in the best number of features 30
6.3.5 Hyperparameter tuning 32
6.3.6 Model comparison 34
6.3.7 Model selection and further evaluation in an independent dataset from Cancer Cell Line Encyclopedia (CCLE) 36
6.3.8 NAPY Classifier subtype-specific gene expression signatures 39
6.3.9 Gene Set Enrichment analysis on the NAPY classifier gene signatures 40
6.4 Gene expression signatures for cancer pathways are differentially expressed in NAPY subtypes 42
6.5 Most genomic alterations are non-mutually exclusive across subtypes 47
6.6 Outcome assessment results 49
6.7 Discussion Summary 52
8. Conclusion 53
9. Supplementary Information 54
10. References 61
11. Acknowledgements 67

1.Background

Small Cell Lung Cancer (SCLC), notorious for its aggressive behavior and rapid progression, poses a great challenge in the field of oncology. It comprises about 15% of all lung cancer cases, with median survival duration of <2 years for patients with early-stage disease and about 1 year for patients with metastatic disease (1).

SCLC has recently undergone subdivision into four distinct molecular subtypes based on the activity of key transcription factors (TFs), namely *ASCL1, NEUROD1, YAP1,* and *POU2F3*, that provide insights into the mechanism and cellular origin of these cancers (2). With support of two decades of molecular studies this subdivision has gained broad acceptance among SCLC investigators and has been referred to in the recent WHO classification of Lung Tumors (3).
While the importance of these transcription factors in SCLC has been widely accepted, the characterization of downstream transcriptional programs linked to each of the TFs is less advanced and supported by only a few studies with limited patient numbers (4) (5) (6).
Moreover, well-defined computational processes for subtype assignments are scarce, with limited clarity on how to translate it to other data. Having a clear and standardized way to classify a new SCLC sample is of utmost need, particularly considering the promising implications highlighted in recent reports regarding subtype assignment for prognosis or therapy selection. This aligns with the evolving treatment paradigm that emphasizes personalized clinical care.

2. Main Goal

Develop an accurate, reproducible, and more sophisticated diagnostic approach for the four SCLC molecular subclasses, facilitating deeper characterization of the subtypes and the exploration of more precise targeted therapeutic strategies.

3. Specific Goals

The goals of this project are:

- Develop a machine learning strategy for SCLC subtype prediction based on real-world data and transcriptional downstream programs associated with the for key transcription factors: *ASCL1, NEUROD1, YAP1, and POU2F3.*

- Validate the biological relevance of our SCLC molecular subclasses through the analysis of the activity of multiple cancer phenomena and pathways, such as proliferation, interferon response, apoptosis, etc.

- Validate the genetic mutations patterns of our SCLC and across subtypes.

- Provide insights into prognosis and treatment-specific survival differences among the four subtypes.

4. Introduction

The main biological concepts regarding SCLC, which includes the subtypes, disease characteristics and treatments are explained in detail below. In addition, the methodology used for data generation and the machine learning techniques are presented.

4.1 Types of pulmonary carcinomas

The list of Lung Tumor types is very extensive, being the major big groups defined by the WHO Classification of Thoracic Tumors in 2021:

- Epithelial tumors,
- Lung Neuroendocrine Neoplasms (NEN),
- Tumors of Ectopic Tissues,
- Mesenchymal Tumors Specific to the Lung and
- Hematolymphoid tumors.

Within each main group, there are specific and more granulated subgroups (3).

In general, the principles of lung tumors classification released by the WHO in 2021 consists mainly in morphological characteristics, supported by immunochemistry and occasionally by molecular techniques (3). The incorporation of immunohistochemistry techniques to support the classification was highlighted by the WHO in 2015, whereas the use of molecular testing was emphasized in the last release of 2021 (3).

4.2 Neuroendocrine cells

Neuroendocrine (NE) cells are rare epithelial cells that, in addition to having an endocrine function, express markers and peptides commonly associated with neurons and the central nervous system. Some of these NE markers commonly used in clinical diagnosis include Chromogranin A (CHGA) - *a secretory protein* -, Synaptophysin (SYP) - *a synaptic vesicle glycoprotein* - and NCAM1 (a.k.a. CD56) - *a cell adhesion molecule* -. Other NE markers are the lineage specific transcription factors *ASCL1, NEUROD1* and *INSM1* which are currently not used in clinical diagnosis (7).

NE cells can be found as either single cells or small clusters of cells dispersed throughout the surface epithelium of different tissues, including the lung, the intestine, and the pancreas, where they will receive a neuronal impulse and perform important endocrine functions. For instance, the beta cells in the pancreas are neuroendocrine cells that secrete insulin (7).

4.3 Lung Neuroendocrine Neoplasms

The neuroendocrine cells located in the lung are called Pulmonary Neuroendocrine Cells (PNECs) and account for only 0.5% of the lung epithelium. PNECs contain secretory granules that produce and secrete bioactive compounds. In this case, the bioactive compounds, hormones, and neuropeptides are known to affect oxygen sensing, pulmonary blood flow and bronchial tonus, and lung immune responses. The bioactive compounds secreted by PNECs include serotonin, calcitonin, calcitonin gene-related peptide (CGRP), gastrin-releasing peptide (GRP), chromogranin A, gamma aminobutyric acid (GABA), synaptophysin, among others (7).

The group of rare tumors that presumably arise from NE cells is known as neuroendocrine neoplasms (NENs). NENs can arise in almost all tissues but they show the highest incidence in the lung and the gastroenteropancreatic (GEP) system (7).

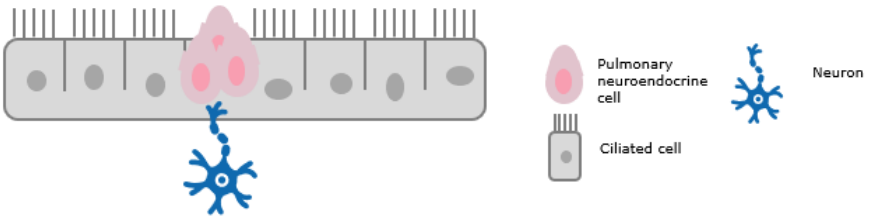

Figure 1. Representation of pulmonary neuroendocrine cells and their connection with neurons.

Lung NENs account for 20-25% of all lung cancers and for 25-30% of NENs from all tissue sites. As is the case for NENs in general, lung NENs comprises two main categories: low grade, well-differentiated Neuroendocrine Tumors (NETs); and high grade, poorly differentiated Neuroendocrine Carcinomas (NECs) (8) (7).

Well-differentiated NETs consist of low-grade Typical Carcinoids (TC) and intermediate-grade Atypical Carcinoids (AC). On the other hand, poorly differentiated NECs are high-grade neoplasms characterized by either small or large cell types, referred to as Small Cell Lung Cancer and Large Cell Neuroendocrine Carcinoma (LCNEC) respectively (8).

The current classification (Figure 2) also retains the categories of combined LCNEC and combined SCLC which may comprise up to 25% of the resected cases. Both categories include either LCNEC or SCLC combined with a Non-Small Cell Carcinoma (NSCC) component, most often adenocarcinoma or squamous cell carcinoma. SCLC may also be combined with LCNEC and is classified under the term "combined SCLC and LCNEC (3).

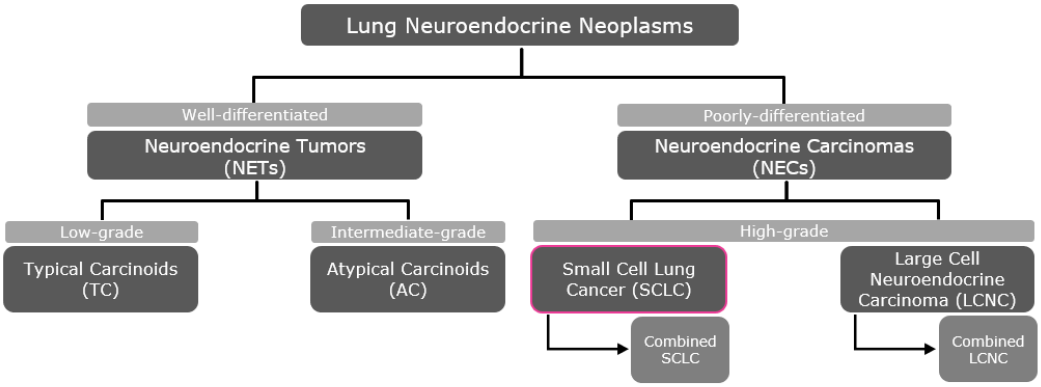

Figure 2. Summary of Lung Neuroendocrine Neoplasms Classification from the WHO release of 2021.

The focus of this project is on Small Cell Lung Cancer, which belongs to the high-grade neoplasms subgroup under the Lung Neuroendocrine Neoplasms category.

Notably, the increase in the biological understanding of lung tumors, along with the incorporation of more advanced techniques into the classification process, has led to the development of more sophisticated and accurate diagnostic approaches, enabling more precise therapeutic strategies. Consequently, a decline in lung cancer mortality has been reported in the United States, correlated with an improvement in patient outcome (3).

However, this is not the case for SCLC, where despite the significant progress in comprehending the molecular biology of SCLC during the past years, no improvement in patient survival has been observed over time. Recently, a new molecular subclassification for SCLC has been proposed, which holds promise for paving the way for a targeted approach to treating SCLC (3).

Nevertheless, significant progress has been made by different research groups in understanding the molecular biology of SCLC during the past years. As a result, a new molecular subclassification for SCLC has been proposed, which holds promise for paving the way for a targeted approach to treating SCLC (3).

4.4. Small Cell Lung Cancer

4.4.1 Disease characteristics

Lung cancer is the leading cause of cancer mortality worldwide. SCLC is a high-grade neuro-endocrine carcinoma arising predominantly in current or former smokers and has an exceptionally poor prognosis (1). SCLC is particularly aggressive, characterized by a predilection for rapid growth, early metastasis and acquired therapeutic resistance (2). SCLC makes up about 15% of lung cancer cases (1).

Respiratory symptoms are commonly seen in patients diagnosed with SCLC. Diagnostic imaging often reveals a lung mass centrally located, along with involvement of thoracic lymph nodes. At the time of initial diagnosis, approximately two-thirds of patients already have metastatic disease. Brain metastases are frequently observed in SCLC, with around 10% of patients presenting with brain involvement during diagnosis, and an additional 40-50% developing brain metastases later. Reflecting its strong tendency to metastasize, SCLC exhibits one of the highest concentrations of circulating tumor cells (CTCs) among solid tumors (1).

4.4.2 Treatment strategies

Even though there is a growing understanding of distinct biological subtypes of SCLC based on the expression profile of transcription factors, the current clinical approach to treating SCLC remains unchanged, as a single disease entity regardless of the subtype. The variability observed in clinical depends solely on the stage of the disease (2).

Early-stage and locally advanced cases are usually treated with radiation and platinum-based chemotherapy together. Patients with metastatic disease receive systemic chemotherapy with or without immunotherapy (PDL1 inhibitor), followed by maintenance immunotherapy (PDL1 inhibitor). Although SCLC exhibits a strong initial responsiveness to cytotoxic treatments, these responses are usually temporary in the majority of patients (1). For recurrent SCLC, approved treatments like topotecan (a topoisomerase I) can be used for second-line treatment and PD1-agonist immunotherapy for third line treatment (2).

Still, patients with early-stage disease have a median survival duration of less than two years, while those with metastatic disease typically have a median survival duration of approximately one year (1).

4.4.3 Tumor morphological and molecular characteristics

As mentioned before for lung tumors and particularly, for the diagnostic of pulmonary NEN subtypes, the classification criteria is mainly based on morphology, with mitoses and presence of necrosis as key factors for the classification (3) (Figure 3). Therefore, the primary method of diagnosing SCLC

involves examining key morphological features of cancer cells within the tumor using stained slides under light microscopy (2).

Immunohistochemistry is frequently used to differentiate SCLC from other diagnoses by employing markers of neuroendocrine differentiation such as chromogranin (CHGA), synaptophysin (SYP), and NCAM[1]. On the other hand, it can also be used for diagnosing SCLC with classic morphology but low expression of neuroendocrine markers, by excluding potential mimics through the staining of protein markers characteristic of those tumors (2).

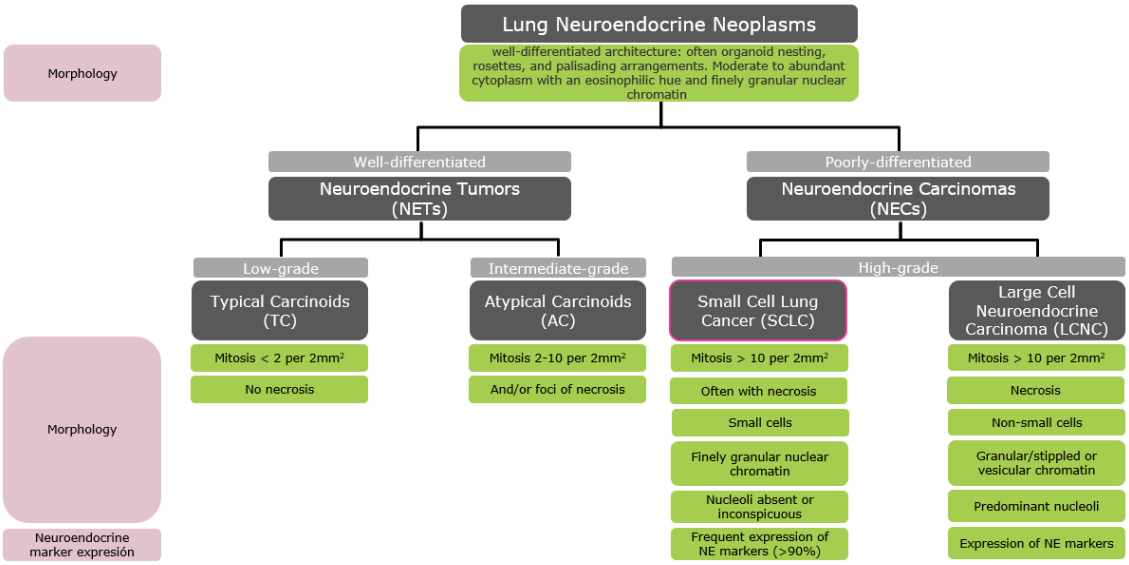

Figure 3. Summary of morphological and molecular characteristics of Lung Neuroendocrine Neoplasms.

The classification of SCLC by WHO recognizes two subtypes: "pure" SCLC (which accounts for approximately 80% of cases) and combined SCLC (comprising around 20% of cases). Combined SCLC has an additional component of non-small-cell carcinoma, which can be of any non-small-cell histological subtype. The most common non-small-cell histological subtype in combination with SCLC are large-cell carcinoma or large-cell neuroendocrine carcinoma (LCNEC). For a diagnosis of pure SCLC, the non-small-cell component must represent less than 10% of the total cell number (1).

Regarding genomic characterization of SCLC, most of the cases exhibit simultaneous inactivation of TP53 and RB1 genes. This dual inactivation of tumor suppressors is distinct from the primary oncogenic drivers of many other solid tumors. However, SCLC is far from being a molecularly homogeneous tumor type. While the SCLC tumor mutational landscape does not seem to define subtypes (1), recent transcriptional data revealed intertumoral heterogeneity and identified at least four distinct subtypes (3) based on the expression of four specific transcription factors: *NEUROD1* (SCLC-N), *ASCL1* (SCLC-A subtype), *POU2F3* (SCLC-P) or *YAP1* (SCLC-Y) (1).

For instance, some of the underlying differences between these subtypes include the degree of neuroendocrine differentiation and variations in metabolism (1).

4.4.4 SCLC subtypes

Over the past 30 years, various research groups have identified and characterized biologically distinct subtypes of SCLC. Initially, these subtypes were characterized through differences in morphology and growth characteristics as well as loss of certain neuroendocrine cell features. Then, through differences in the degree of neuroendocrine differentiation and, more recently, through differential expression of different transcription regulators (2). Consequently, a variety of

terminologies have been applied over the years to identify the different subtypes, until *Rudin et. at*. proposed in 2019 a working nomenclature for SCLC distinct subtypes, defined by the relative expression of four key transcription factors namely *NEUROD1*, *ASCL1*, *POU2F3* and *YAP1* referred to as the NAPY classification (2) (9). Since then, many research groups have used this nomenclature to expand and deepen the understanding of the subtypes of SCLC. Please refer to Figure 4, generated from Rudin *et. al* (2019) summary and further on research studies.

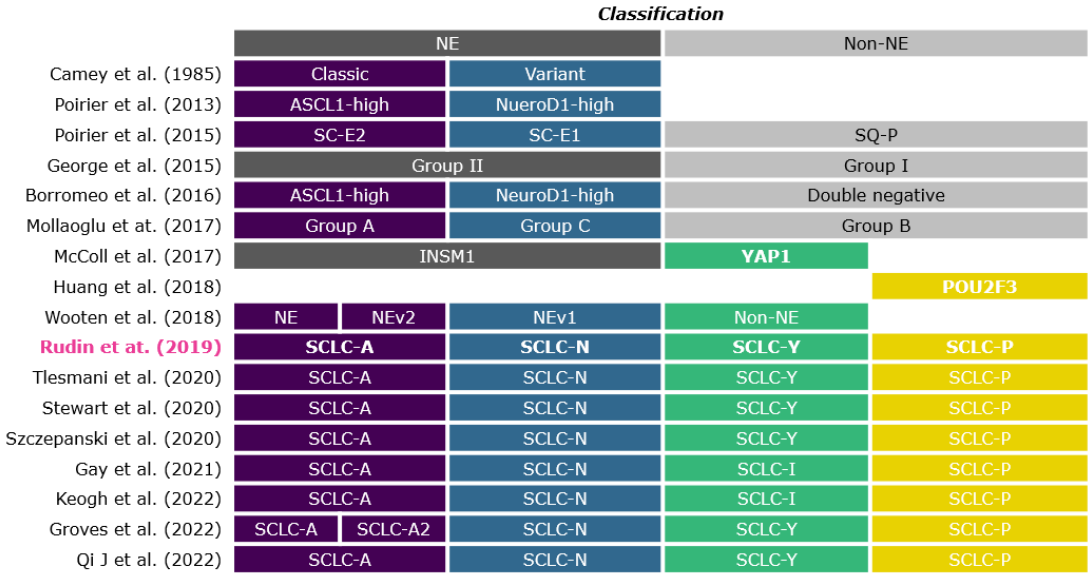

Figure 4. Summary of the molecular classification of SCLC over time.

Evidence on the SCLC subtypes has come both from in-vitro and in-vivo samples, such as human tumors, xenografts, cell lines and genetically engineered mouse models (GEM) (2).

First, a “variant” form of the “classic” and more common SCLC cell lines was described, based on specific growing characteristics (2). The variant subtype was frequently associated with loss or decreased expression of neuroendocrine features, often loss of *ASCL1* transcription factor and expression of *NEUROD1* transcription factor or, loss of both transcription factors (9). These characteristics were also noted in primary SCLC tumors from patients, including autopsy specimens (2). In addition, the *variant* cell lines appeared to be from post-therapy tumors that had recurred, to be more radioresistant than the classic lines and presented more frequent amplifications in the *MYC* family genes than the classic cell lines (9).

In 2012 and 2015, the firsts genomic profiling studies (WGS and WES) of human SCLC were published. Despite the comprehensiveness of the studies, they did not share light on the classification of SCLC subtypes, mainly because recurrent genomic alterations did not consistently co-occur together or exclusively. In fact, relative genetic homogeneity was reported, with almost universal inactivation of *TP53* and *RB1* tumor suppressor genes (2).

Epigenetic and gene expression studies, however, demonstrated molecular heterogeneity in SCLC cell lines and primary tumors, indicating the existence of at least four subtypes. Within these subtypes, two were characterized by high expression levels of either *ASCL1* or *NEUROD1*, while the other two, by low levels of both *ASCL1* and *NEUROD1* (2). Additional studies identified that *ASCL1* and *NEUROD1* activate non-overlapping set of genes implicated in SCLC biology, suggesting that tumors with high levels of *ASCL1* and *NEUROD1* represent distinct subtypes of SCLC (2). Some studies suggest that the *NEUROD1*-high subtype represents an intermediate phenotype with properties that lie between the high and low neuroendocrine phenotypes (9).

Regarding the subtype characterized by low expression levels of both *ASCL1* and *NEUROD1* transcription factors, expression profiling studies have reported differential expression of *YAP1*, a transcriptional regulator activated by the HIPPO growth signaling pathway, compared to the SCLC lines in the *ASCL1*-high or *NEUROD1*-high subtypes. On the contrary, *INSM1* has been reported to be preferentially expressed in neuroendocrine phenotype. However, it remains unverified whether *YAP1* functions as a transcriptional driver specific to this subtype or serves as a correlate associated with this particular subtype (2).

Furthermore, another study has reported *POU2F3* transcription factor as differentially expressed in SCLC cell lines lacking high levels of expression of both *ASCL1* and *NEUROD1*. This transcription factor was found to have specific downstream genes linked to it that were previously identified as overexpressed in non-neuroendocrine SCLC. Even though *POU2F3*-positive tumors express variable levels of *YAP1*, the relative expression of *YAP1* and *POU2F3* defines distinct populations of *YAP1*-high and *POU2F3*-high tumors, defining a third and fourth subtype of SCLC from the non-neuroendocrine subtype (2).

4.4.5 Cell of origin

SCLC has long been assumed to initiate in neuroendocrine lung epithelial cells. Nevertheless, some findings suggest that, in addition to a subset of neuroendocrine cells, other lung epithelial cells may serve as cells of origin (1). For instance, some studies have proposed the possibility that *POU2F3*-high subtype might originate from tuft cells, which are a rare chemosensory cell type in the pulmonary epithelium. This hypothesis is based on the selective expression of *POU2F3* in tuft cells and the similarity of expression profiles. However, the authors have emphasized that further investigations are needed to validate this hypothesis (2).

4.4.6 Intratumor heterogeneity

Some studies using single-cell RNA sequencing data in mouse models propose that different molecular subtypes may correspond to different stages of SCLC progression, where tumors initially emerge in an *ASCL1*-high state and then transition towards a non-neuroendocrine state (7). For instance, it has been postulated that this dynamic evolution of SCLC subtypes could be driven by *MYC*. In neuroendocrine cells, *MYC* could activate Notch signaling to induce a fate shift in SCLC from *ASCL1*+ to *NEUROD1*+ to *YAP1*+ states. *MYC* alternatively promotes *POU2F3*+ tumors from distinct cell types (10) (11). Furthermore, it has been observed that individual tumors may consist of cells belonging to different subtypes, highlighting intratumor heterogeneity in SCLC (7).

4.4.7 Method for subtype classification

Lung cancer-specific up and down regulated neuroendocrine signatures have been published (9), which could be used to distinguish neuroendocrine phenotypes of SCLC from non-neuroendocrine phenotypes, but these do not have the granularity to classify on the four SCLC subtypes described.

When the sample specimen is available in the wet-lab, some research groups have used immunohistochemical staining to detect the protein expression levels of *NEUROD1*, *ASCL1*, *POU2F3* and *YAP1* and classify on the four subtypes using a scoring system (H-score) based on the staining intensity and the percentage of positive cells of the indicated protein (12) (13) (14). However, in some immunohistochemistry analysis, due to the low protein expression level of *YAP1* in SCLC tissues, SCLC-Y could not be observed. As a consequence, some research groups have proposed to refer to this subtype as “triple-negative” or “SCLC-I”, accounting for the low expression of all three transcription factors (*NEUROD1- ASCL1-POU2F3*) and the reported inflamed characteristics of SCLC-Y subtype (12) (13) (15).

On the other hand, when gene expression is available, SCLC samples are classified as SCLC-N, SCLC-A, SCLC-P or SCLC-Y, based on the expression level of the transcription regulator

(*NEUROD1*, *ASCL1*, *POU2F3* or *YAP1*) with the greatest relative overall expression, following the NAPY classification (2) (16).

Nevertheless, there is yet no guideline from the WHO on a fixed classification based on the expression of these 4 transcription factors, nor is there an harmonized classification algorithm that could be used for SCLC sample classification.

Using the expression of subtype-specific downstream programs associated to the four master transcription factors *NEUROD1*, *ASCL1*, *POU2F3* and *YAP1* could provide an additional layer of robustness to the subtyping procedure, compared to classifying solely based on the four transcription factors. In this regard, many genes have been reported to be preferentially expressed in one or another subtype (17) (18), but very few studies have attempted to classify SCLC subtypes based on subtype-specific downstream programs associated with these four key transcription factors (4).

4.4.8 Therapeutic Implications

As mentioned before, although most SCLC initially respond well to chemotherapy, the development of resistance is essentially universal, and patients are rarely cured (17).

Even though the addition of immunotherapy in the treatment of SCLC is reaching a relatively mature stage, there is still much to explore regarding more personalized therapeutic strategies that account for the distinct molecular subtypes and heterogeneity observed in SCLC (19).

In this regard, there is a growing recognition that the distinct molecular subtypes of SCLC may display variability in their response to different therapies, when evaluated in preclinical models of the disease (1) (19).

Table 1 shows a summary of some therapeutic approaches that have been reported to be potentially beneficial to a subset of patients based on the molecular characteristics of the disease.

Treatment		**SCLC-N**	**SCLC-A**	**SCLC-P**	**SCLC-Y**	***Reference***
Growth and survival signaling pathways - apoptosis	BCL2i[1]		X			*Poirier et al (2020), Schwendenwein et al. (2021), Yatabe et al (2020)*
Antitumor immunity	DLL3i[2]		X			*Poirier et al (2020), Schwendenwein et al. (2021)*
Epigenetic modifier	LSD1i[3]	X	X			*Poirier et al (2020), Schwendenwein et al. (2021), Yatabe et al (2020)*
DNA Damage Response (DDR) proteins	AURKA/Bi[4]	X		X	X	*Poirier et al (2020), Schwendenwein et al. (2021)*
DNA Damage Response (DDR) proteins	CHK1[5]	X		X	X	*Poirier et al (2020)*
Purine biosynthetic pathway	IMPDH6	X		X	X	*Poirier et al (2020)*
Growth and survival signaling pathways	IGF1Ri[7]			X		*Poirier et al (2020) , Schwendenwein et al. (2021), Yatabe et al (2020)*
Antitumor immunity	IOi[8]				X	*Poirier et al (2020), Schwendenwein et al. (2021)*
DNA Damage Response (DDR) proteins	PARPi[9]			X		*Schwendenwein et al. (2021)*
Epigenetic modifier	HDACi[10]		X			*Poirier et al (2020), Schwendenwein et al. (2021)*
Growth and survival signaling pathways	mTORi[11]				X	*Schwendenwein et al. (2021)*

[1] BCL-2, B-cell lymphoma ; [2] DLL3, Delta-like canonical Notch ligand 3; [3] LSD1, Lysine-specific Demethylase 1a; [4] AURKA/B, Aurora Kinase A/B; [5] CHK1, Checkpoint kinase 1; [6] IMPDH: Inosine monophosphate dehydrogenase ; [7] IGFR1, Insulin like Growth Factor 1 Receptor ; [8] IO, Immunotherapy ; [9] PARP, Poly (ADP-ribose) polymerase ; [10] HDAC, Histone deacetylase ; [11] mTORi, Mammalian target of rapamycin

Table 1. Summary of proposed therapeutic strategies with potential benefit for a molecular subset of SCLC.

For instance, it is expected that SCLC-A subtype would exhibit a positive response to antibody drug conjugates targeting *DLL3*, as a result of the direct interaction between *DLL3* and *ASCL1* at the transcriptional level in tumor cells with inactive Notch signaling. Similar applies to *BCL2* inhibitors, considering *BCL2* is a direct transcriptional target of ASCL1 (17).

On the other hand, xenograft models of *ASCL1*^{low} SCLC cell lines, and GEMMs of *ASCL1*^{low}/ *MYC*^{high} tumors were reported to be sensitive to *IMPDH1*/2 inhibitors (19).

Other studies suggest that IGF-1R inhibitors and PARP inhibitors could potentially be targeted therapeutic agents for patients within the SCLC-P subtype (17).

In addition, since *YAP1* has been reported to upregulate PD-L1 transcripts and induce an immunosuppressive tumor microenvironment, it has been suggested that SCLC-Y may be preferentially sensitive to Immunocheckpoints inhibitors (17).

Despite the promising of these targeted therapies, the results seen so far in clinical studies of these therapies applied to SCLC have been disappointing. The rationale that has been attributed to these findings lies behind the heterogeneity of SCLC, since patients are still enrolled in clinical trials irrespective of their molecular background (17).

The above emphasizes the need for a standardized way of molecularly classifying SCLC subtypes to guide treatment for targeted therapies (20).

4.5. Gene expression overview

4.5.1 The foundations of gene expression

The fundamentals of gene expression are governed by the central dogma of molecular biology, which involves the transcription of the genetic information stored in the DNA into RNA molecules, and, if applicable, finally translation into proteins (21).

The transcription of a subset of genes into RNA, namely the transcriptome, plays a crucial role in determining a cell's identity and regulating its biological activities (21).

The transcriptome is a complex entity consisting of various types of coding and noncoding RNA species. While messenger RNA (mRNA) has historically been the main focus due to its role in encoding proteins, noncoding RNAs (ncRNAs) are also functional and diverse. Traditional ncRNAs perform essential cellular functions, like ribosomal RNA (rRNA) and transfer RNA (tRNA) involved in translation, small nuclear RNA (snRNA) involved in splicing, and small nucleolar RNA (snoRNA) in the modification of rRNAs, among others. Nevertheless, new classes of ncRNAs have also emerged recently, such as microRNA (miRNA) and piwi-interacting RNA (piRNA), both of which have been reported to regulate gene expression at the posttranscriptional level (21).

4.5.2 Next Generation Sequencing platforms

RNA sequencing (RNA-Seq) is a powerful technique that uses high-throughput sequencing methods, providing higher coverage and greater resolution of the dynamic nature of the transcriptome, over older methods like Sanger sequencing-based approaches and hybridization-based microarrays. The basis of RNA-Seq experiment consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform (21).

Currently, there are multiple commercially available next-generation sequencing (NGS) platforms accessible, several of which use a sequencing-by-synthesis method. These platforms are typically classified based on the length of the read sequenced. For example, Ensemble-based platforms such

as Ion Torrent facilitate the sequencing of many identical copies of short DNA molecules, while Single Molecule Real Time (SMART) based platforms like PacBio allow for the sequencing of a single, much longer DNA molecule. The choice of sequencing technique and platform has implications for downstream analysis and interpretation of the sequencing data. Illumina, which has been a dominant player in the sequencing industry, employs an ensemble-based sequencing-by-synthesis approach, where DNA molecules are amplified while attached to a glass flowcell, and synthesis utilizes reversible-terminator nucleotides labeled with fluorescence (21).

4.5.3 RNA-Seq Data Analysis Workflow

The traditional pipeline for RNA-Seq data involves obtaining the reads from a next-generation sequencing (NGS) platform, aligning these reads to a known reference genome, and quantifying gene expression (21) .The result is a set of the detected genes with its from reads estimated expression, which constitutes the starting point for several downstream analyses (e.g. Gene expression signatures).

4.5.4 Gene expression signatures

Gene expression signatures are modules of coexpressed genes that describe functional aspects and characteristics of a cell. These signatures can describe cell-type composition, signaling pathways activities, activities of cellular processes like proliferation or interferon response, or can even inform on the cell-of-origin (22).

In this regard, gene expression signatures have been widely used not only to characterize tumor-specific gene expression phenotype, but also important cancer phenomena like oncogenic signaling pathway activities, or immune cell infiltration into tumor tissues. Today, the investigation of expression signatures has become a standard component in the elucidation of tumor biology, but also in clinical use, to inform on patient prognosis or predict therapeutic efficacy (22).

4.6 Machine Learning Overview

Machine learning techniques encompass a broad range of algorithms and methods used to enable computers to learn from data and make predictions or classifications without being explicitly programmed for every task.

4.6.1 Supervised Learning

In machine learning, supervised learning involves fitting a model to predict a *response variable* based on predictor measurements, aiming to accurately predict future responses or understand the relationship between predictors and responses. This concept opposes that of unsupervised learning, where there is no associated *response variable* for each observation, which means working without guidance or supervision, hence the term "unsupervised" (23).

4.6.2 Classification problems

Response variables can be classified as either quantitative or qualitative (also known as categorical). Quantitative variables are characterized by numerical values. On the other hand, qualitative variables assume values within distinct classes or categories, such as a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia). Problems involving a quantitative response are often referred to as regression problems, while those with a qualitative response are commonly termed classification problems (23).

The simplest setting in a classification problem is a binary (two-class) classification. However, classification problems usually extend to k classes (23). Some algorithms natively support classification tasks with more than two classes (e.g. nearest centroid, k-nearest neighbors, random forest), while others were designed for binary classification (e.g. support vector machine). The extension of the support vector machine algorithm to a multiclass classification of k classes, with k

> 2, can be achieved with a 'one-against-one' approach, in which k(k − 1)/2 binary classifiers are trained. The appropriate class is assigned by a voting scheme (24). In case of prediction conflicts (i.e., equal number of votes for multiple classes), the distance of the data points to the decision boundary or hyperplane can serve as a measure of the prediction's robustness, aiding in the determination of the final predicted class.

4.6.3 Nearest Centroid

The Nearest Centroid Classifier (NCC) is a centroid-based classification method. In this method, the training data is used to calculate the centroid of each class label, which is determined as the mean of all data points within that class. Subsequently, each data point from the testing set is allocated to the nearest class label, based on minimizing the distance between the data point and the centroid of the class. The advantage of NCC is that the method does not have parameters, so the results obtained do not depend on parameter configuration, but purely on the distance between data (25).

4.6.4 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classifier involves selecting a positive integer K, and then for each test observation x_o, identifying the K closest neighbors to x_o in the training data (represented by N_0). The classifier then estimates the conditional probability for a given class *j* by calculating the fraction of points in N_0 with response values equal to *j*, as below:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Where Y represents the *response variable,* and X the observation.

In practice, K is treated as a tuning parameter that is generally chosen via cross-validation.
Finally, KNN classifies the test observation x_o to the class with the largest conditional probability (23).

4.6.5 Support Vector Machine

The Support Vector Machine (SVM) is a further extension of the Support Vector Classifier in order to accommodate non-linear class boundaries. Or, in other words, the Support Vector Classifier is a specific case of the SVM where the classes are separable by linear boundaries (23).
The goal of the Support Vector Classifier is to construct a hyperplane that separates the training observations according to their class label, and then classify the test observations depending on which side of the hyperplane it is allocated using its feature measurements (23).

The equation of an hyperplane in the p-dimensional space is:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

And has the following property, where y_i is the label of the observation:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$$

The optimal separating hyperplane is chosen from the infinite possible hyperplanes based on the maximal margin hyperplane. The margin (M) is the perpendicular distance from each training observation to a given separating hyperplane. Intuitively, the maximal margin hyperplane is the separating hyperplane for which the margin is largest (23). Maximizing the margin ensures that each observation is on the correct side of the hyperplane and at least a distance M from the hyperplane.

The support vectors are the observations that define and “support” this maximal margin hyperplane, in the sense that if these points were moved slightly, then the maximal hyperplane would move as well. (23) Therefore, in order to have greater robustness to individual observations (avoid overfitting on the training data), and do a better job in classifying the remaining observations, the method accepts the misclassification of a few observations.

Consequently, constructing the maximal margin hyperplane from the set of n training observations $x_1,...,x_n$ and their associated class labels $y_1,...,y_n$, corresponds to the solution to the optimization problem below:

$$\begin{aligned}
&\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M \\
&\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \\
&y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i), \\
&\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,
\end{aligned}$$

where *M* is the width of the margin, $\epsilon_1,...,\epsilon_n$ are slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane; and C is a nonnegative parameter representing the number and severity of the violations to the margin (and to the hyperplane) that could be tolerated and therefore, known as the Cost (23).

A value of $\epsilon_i = 0$ for the slack variable means that the ith observation is on the correct side of the margin, whereas a value between 0 and 1 means that the ith observation has violated the margin, and values of $\epsilon_i > 1$ means that it is on the wrong side of the hyperplane (23).

On the other hand, a value of Cost = 0, implies that there is no budget for violations to the margin (all observations must be on the correct side of the margin: $\epsilon_i = \ldots = \epsilon_n = 0$). As the budget C increases, there is more tolerance to violations to the margin, and so the margin widens (models are potentially more biased but may have lower variance). Conversely, as C decreases, there is less tolerance to violations of the margin and so the margin narrows (models are potentially less biased but have high variance). In practice, C is treated as a tuning parameter that is generally chosen via cross-validation (23).

Computationally speaking, the solution to the above optimization problem is reduced to the inner products of the observations, which can be represented as:

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle,$$

where S in the collection of indices of the support vector points.

When there is a non-linear relationship between the predictors and the outcome, the feature space can be enlarged by using functions of the predictors (ex: quadratic function). These transformation functions are referred to as kernels.

In these cases, the inner product in the above formula is replaced by the kernel.

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i).$$

Where the kernel represents the function that quantifies the similarity of two observations.

4.6.6 Random Forest

A Random Forest (RF) classifier is built on a number of decision trees on bootstrapped training samples.

Decision Trees are a set of splitting rules used to segment the predictor space into a number of simple regions. In order to make a prediction for a given observation, typically the mean *response value* for the training observations in the region to which it belongs is used (23).

The process of building a regression tree has mainly two steps:

- Dividing the predictor space (X_1, X_2, . . . , X_p) into *J* distinct and non-overlapping regions (R_1,R_2, . . . ,R_J)
- For every observation that falls into the region R_j, the same prediction is made, which is the mean of the response values for the training observations in R_j.

The goal is to find regions R_1, . . . , R_J that minimize the Residual Sum of Squares (RSS), given by:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

where $\hat{y}_{Rj}$ is the mean response for the training observations within the j_{th} box.

The above is done in a top-down, greedy approach that is known as recursive binary splitting.

That is, at each step, all predictors X_1, . . . ,X_p, and all possible values of the cut point s for each of the predictors is considered, and then the predictor and cut point are chosen such that the resulting tree has the lowest RSS.

$$R_1(j, s) = \{X | X_j < s\} \text{ and } R_2(j, s) = \{X | X_j \geq s\},$$

j y *s* must minimize the following equation::

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

At each step of the tree-building process, the best split is made at that particular step.

In classification trees, instead of predicting the response of an observation as the mean response of the training observations that belong to the same terminal node of the tree (as in regression trees), the prediction is done based on the *most commonly occurring class* among training observations in the region to which the observation belongs (23).

In the classification setting, the *classification error rate* is used instead of the RSS, which is the fraction of the training observations in the region that do not belong to the most common class. When building a classification tree, the quality of a particular split is typically evaluated with the Gini index or the entropy, which are related to the classification error rate, and represent better representative the node purity (23).

When building the decision trees for the Random Forest, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors. A fresh sample of *m* predictors is taken at each split (23).

Complex trees may produce good predictions on the training set, but are likely to overfit the data, leading to poor test set performance. On the contrary, a smaller tree with fewer splits (that is, fewer regions $R_1, \ldots, R_J$) might lead to lower variance and better interpretation at the cost of a little bias (23).

4.6.7 Performance Metrics

Classification algorithms are evaluated based on performance. Multiple measures for assessing classification quality can be used depending on the situation. Some of the most commonly used measures are accuracy, sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value (26).

A confusion matrix is among the most popular tools utilized for validating classification performance, as it serves as the foundation for calculating the classification performance metrics. It consists of a table presenting the various combinations of predicted and actual values associated with the classes. The diagonal of the confusion matrix includes the correctly classified samples, while the off-diagonal cells represent the errors (misclassified samples) (26), refer to Figure 5.

	Predicted class A *(positive)*	**Predicted Class B** *(negative)*
Actual class A *(positive)*	**True Positive (TP)**	**False Negative (FN)** *Type II Error*
Actual class B *(negative)*	**False Positive (FP)** *Type I Error*	**True Negative (TN)**

Figure 5. Confusion matrix example (binary setting).

In multiclass classification, performance can be measured using either macro or micro averaging. With macro averaging, the metric is calculated independently for each class (as a binary setting), and then the average of these values is computed. This approach ensures equal importance for all classes, regardless of their size. On the other hand, micro averaging considers all classes together, where each correctly classified sample contributes to the overall count of correct classifications.

Table 2 describes the formula and interpretation of the different performance metrics, generated from the confusion matrix, for both binary and multiclass problems (26).

Metric	Binary problem	Multiclass problem *we define the sum of all samples as s*
Accuracy *It is the ratio of the number of accurate predictions to the total number of predictions.*	$= \frac{TP + TN}{TP + FP + FN + TN}$	$= \frac{\sum_{i=1}^{c} TP_i}{s}$
Precision (Pos Pred Value) *Determines how many samples, out of all those classified as positive, are samples of the positive class.*	$= \frac{TP}{TP + FP}$	*Macro:* $= \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{TP_i + FP_i}$ *Micro:* $= \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + \sum_{i=1}^{c} FP_i}$
Recall (Sensitivity) *Determines how many samples belonging to a positive class were classified as positive by the classifier.*	$= \frac{TP}{TP + FN}$	*Macro:* $= \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{TP_i + FN_i}$ *Micro:* $= \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} TP_i + \sum_{i=1}^{c} FN_i}$
Specificity *Determine the number of samples belonging to a negative class that have been classified as negative by the classifier.*	$= \frac{TN}{FP + TN}$	*Macro:* $= \frac{1}{c} \sum_{i=1}^{c} \frac{TN_i}{TN_i + FP_i}$ *Micro:* $= \frac{\sum_{i=1}^{c} TN_i}{\sum_{i=1}^{c} TN_i + \sum_{i=1}^{c} FP_i}$
Neg Pred Value *Determines the number of samples correctly classified as a negative class, from all samples that have been classified as negative by the classifier.*	$= \frac{TN}{FN + TN}$	*Macro:* $= \frac{1}{c} \sum_{i=1}^{c} \frac{TN_i}{TN_i + FN_i}$ *Micro:* $= \frac{\sum_{i=1}^{c} TN_i}{\sum_{i=1}^{c} TN_i + \sum_{i=1}^{c} FN_i}$
F1	$= 2 \times \frac{precision\_binary \times recall\_binary}{precision\_binary + recall\_binary}$	*Macro:* $= 2 \times \frac{macro\_precision \times macro\_recall}{macro\_precision + macro\_recall}$ *Micro:* $= 2 \times \frac{micro\_precision \times micro\_recall}{micro\_precision + micro\_recall}$
Balanced Accuracy *Calculates balanced accuracy for, where significant differences in the class size are observed.*	$= \frac{recall\_binary + specificity\_binary}{2}$	$= \frac{1}{c} \sum_{i=1}^{c} \frac{TP_i}{TP_i + FN_i}$

Table 2. Formula and interpretation of classification performance metrics.

In our multiclass setting, macro-averages of the performance metrics across classes are being used, in order to get a unique performance value per metric and facilitate model comparison.

Macro-averaging treats all classes equally by calculating the metric independently for each class and then averaging the results. This ensures that each class contributes equally to the overall metric, regardless of class imbalance, making it sensitive to the performance of rare classes. In addition, Macro-averaging provides a straightforward interpretation, as it represents the average score across all classes. Since our interest is in the overall performance across all classes, macro-average provides a good measure of the generalization performance of the model.

4.6.8 Cross Validation

The cross-validation approach (Figure 6) involves randomly dividing the set of observations into k groups (or folds) of approximately equal size. The model is then fitted using k-1 folds and validated using the remaining fold. This process is repeated k times, each time using a different fold as the validation set, resulting in k performance estimates (e.g., k accuracy estimates). The k-fold cross-validation estimate is obtained by averaging these values (23).

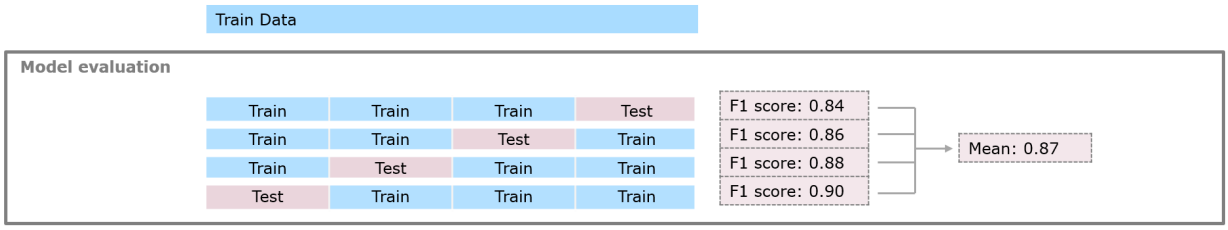

Figure 6. Example of a 4-fold cross validation. The numerical values are included for explanatory purposes.

Cross Validation is fundamental to estimate the generalization performance of a classifier for a given dataset, and it has been extended in multiple ways (like Nested Cross Validation *-nCV-*) to incorporate hyperparameter tuning (26).

The nCV approach enables both the estimation of performance and model selection (choosing best hyperparameter) without resulting in an overly optimistic generalization error due to overfitting during model selection, which can occur with a simple cross-validation approach (Cawley et al 2010). Briefly, the process of nCV (Figure 7) involves nesting two cross-validation procedures: an outer cross-validation loop where the model is evaluated; and an inner cross-validation loop from each training subset to evaluate hyperparameters and select the best hyperparameter for that run. Once the best hyperparameter is chosen, it is used to train in the complete training subset of that run and test in its corresponding outer test fold (26).

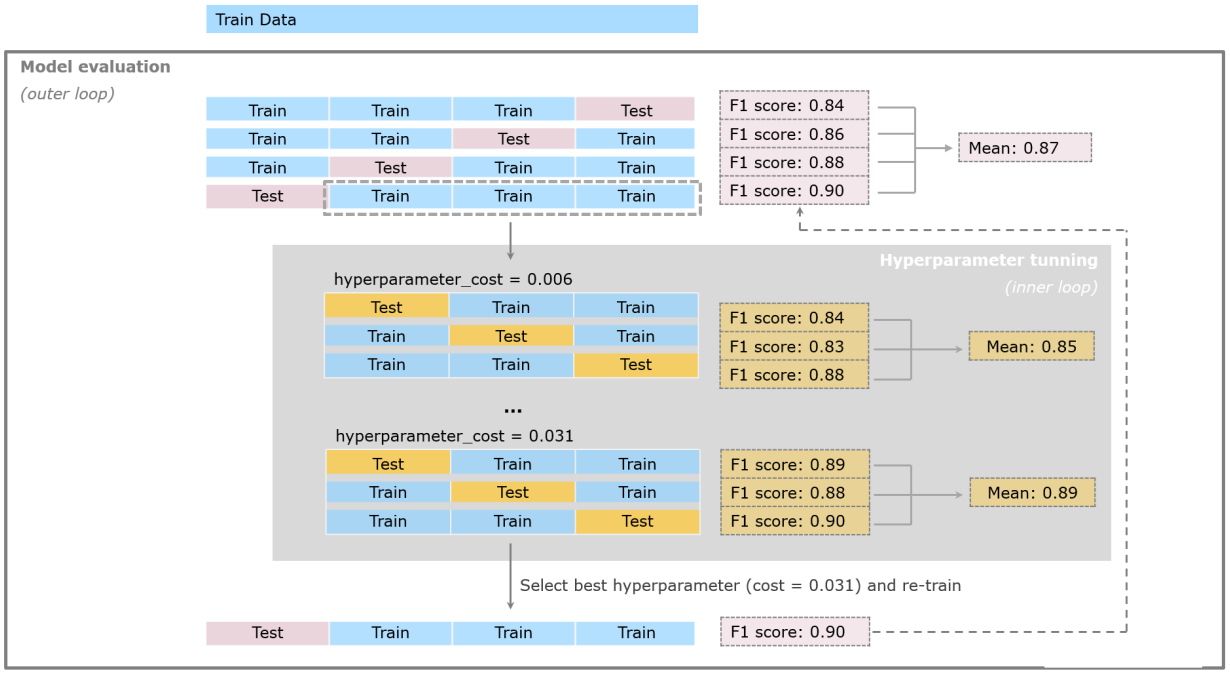

Figure 7. Example of nCV consisting of 4-fold cross-validation in the outer loop and 3-fold cross-validation in the inner loop. The numerical values are included for explanatory purposes.

5. Datasets and methods

5.1. Datasets

5.1.1 Tempus Dataset

Merck has in-licensed a multi-omics real-world evidence database from the company Tempus AI, Inc., containing clinical and molecular data from several thousand cancer patients, including patients with SCLC. The data is subject to controlled access for privacy and proprietary reasons.

The clinical and molecular data has been shared to Merck in an already processed form. Tempus internal assays, including RNA-Seq and DNA-Seq assays were analytically validated to provide high-quality data.

5.1.1.1 Transcriptomic data

This study uses de-identified transcriptomic data (RNA-seq) from SCLC tumor samples from the Tempus Database sequenced by the Tempus xR assay. Briefly, Tempus xR is a whole-transcriptome capture next-generation sequencing assay that quantifies transcript- and gene-level expression, and identifies transcriptional evidence of chromosomal rearrangements resulting in the expression of fusion RNA species. Expression quantification is provided to us both as raw counts and TMP values.

SCLC samples included in our analysis have a minimum of 15 million total reads and no more than 80 million. Additionally, we filtered out samples with diverging count distributions, having a median log fold change distribution of reads greater than 0.3 in absolute value with respect to the median of all log fold change distributions after three iterations, each using a different random reference sample.

For our analysis log2 transformed TPM batch corrected + 1 was utilized. Batch correction enables the removal of potential influences by other technical factors (experimental assay, tumor content fraction, subject ethnicity, etc.).

5.1.1.2 Genomic data

De-identified targeted DNA sequencing data were analyzed for the same SCLCs patients from the Tempus Database. Tumors were profiled using the Tempus xT assay, a next-generation sequencing DNA-seq panel capturing 598 or 648 genes depending on assay version. For the genomic characterization of the SCLC tumor samples from Tempus, the comparison of genomic alterations across molecular subtypes was performed on single-nucleotide variants and copy number variations.

5.1.2 Cancer Cell Line Encyclopedia (CCLE) Dataset

Transcriptomic data obtained from 48 human cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE), publicly downloaded from Cancer Dependency Map project (DepMap). For these cell lines, gene expression TPM values of the protein coding genes were inferred from RNA-seq data using the RSEM tool and reported after log2 transformation, using a pseudo-count of 1; as log2(TPM+1). (27) Refer to the cell lines identification of the Supplementary Data Table 1.

5.2. Subtype Assignment based on Transcriptomic data

5.2.1 Tempus Labeling

Subtype labeling has been done following the NAPY classification (*NEUROD1*-high, *ASCL1*-high, *POU2F3*-high, *YAP1*-high). The subtype-class was assigned based on the highest, z-scaled gene expression, among the four key transcription factors (*NEUROD1, ASCL1, POU2F3, YAP1*), and requiring the highest expression to be at least 0.3 units greater than the rest, and not all-negative expressions. Samples not satisfying the above requirements were not assigned to any specific subtype.

Other more stringent thresholds (0.4, 0.5) for the minimum difference in expression of the 4 transcription factors were assessed for class assignment. However, fewer samples could be classified into the 4 established classes, and no clear improvement in classifier performance was observed.

5.2.2 CCLE Labeling

Subtype labels were taken from Rudin *et al.* (2019) supplementary data, which were assigned to one of four subtypes based on the transcription regulator (*NEUROD1, ASCL1, POU2F3, YAP1*) with the greatest relative overall expression (2) (28). The dataset consists of 11 *NEUROD1*-high, 26 *ASCL1*-high, 4 *POU2F3*-high and 7 *YAP1*-high SCLC cell lines.

5.3. Machine Learning Classifier Development

The pipeline begins by randomly splitting the Tempus expression data in an 80:20 ratio. Stratification was used to ensure that the proportion of each class is maintained during the split. On the 80% of the patient records the prediction accuracy of several machine learning models (nearest centroid, k-nearest neighbors, linear support vector machine, and random forest) was robustly estimate through a nCV approach, enabling both the optimization of hyperparameters and model selection. The remaining 20% of the patient records were left untouched and not used neither in feature selection nor in classifier training, but solely used for benchmarking of the final NAPY classifier.

The nCV setting built on the 80% of the patient's records (training data), involved an outer 4-fold CV executed on 5 random data splits resulting in 20 different training-testing runs for the outer cross-validation, and an inner 3-fold CV (Figure 8).

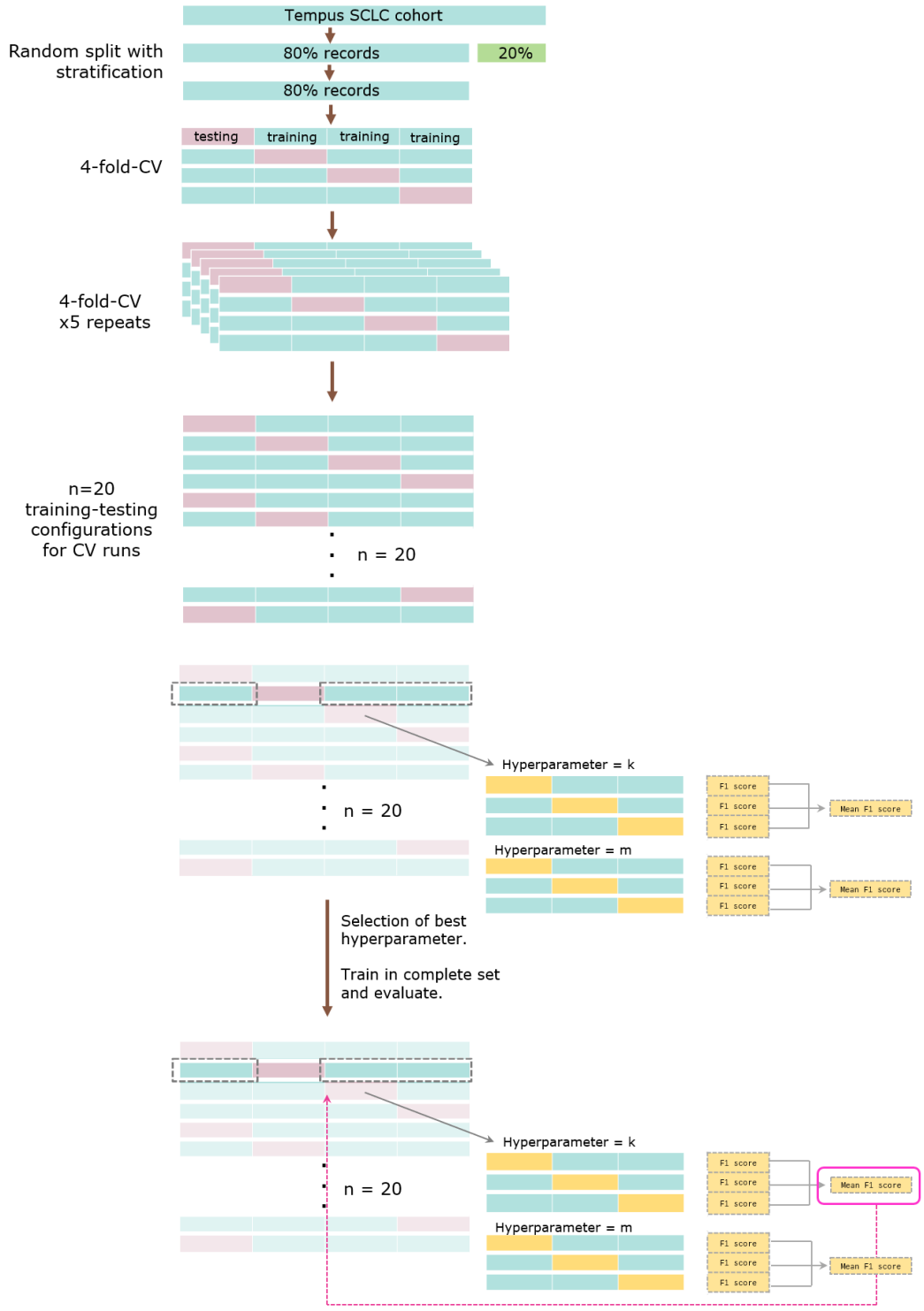

Figure 8. Our approach involved a nested cross-validation setting built on 80% of the patient records, employing a 4-fold cross-validation in the outer loop, executed on 5 random data splits, resulting in 20 outer cross-validation runs, and a 3-fold cross-validation in the inner loop. The remaining 20% of patient records was reserved for testing the final NAPY classifier.

5.3.1 Feature Selection

The predictor genes for the model were chosen from the universe of genes, to ensure the representation of downstream programs associated with each of the four transcription factors: *NEUROD1, ASCL1, POU2F3* and *YAP1*, but excluding these four genes that had provided the class labels. These programs were built based on the top genes with strongest and exclusive association with one of the transcription factors as determined by the Pearson correlation between the entire set of genes and the key TFs, in the training part of each the 20 outer CV runs, to avoid potential leak of information from the testing part. This procedure resulted in a grid of 20 slightly different group of features (Figure 9).

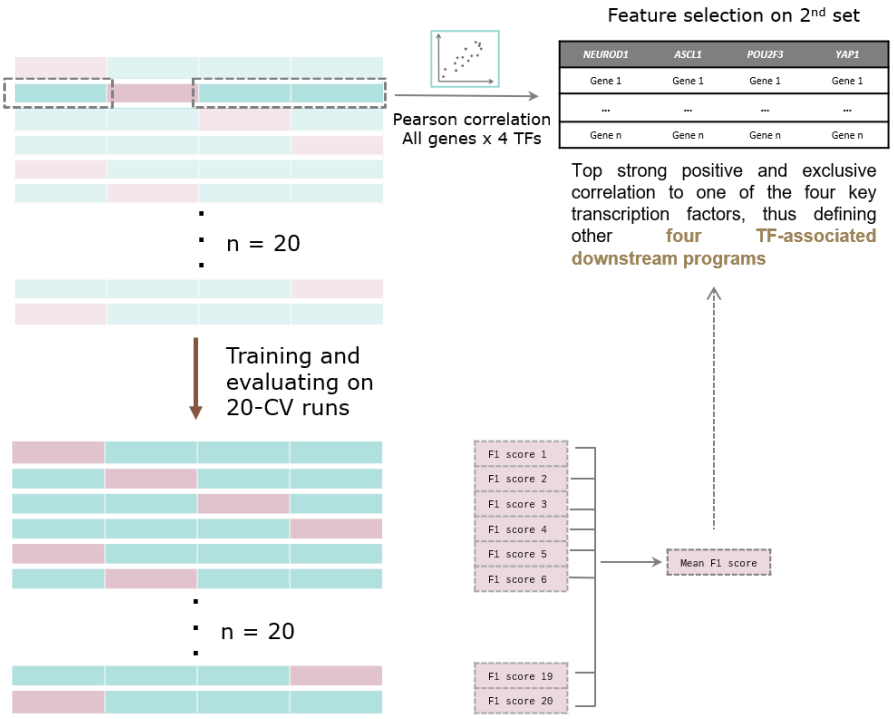

Figure 9: Top genes selected based on a strong positive, and exclusive association with one of the TFs, measured by Pearson correlation in the training part (k-1 folds) of each of the 20 CV configurations.

We assessed different TF-downstream program lengths by including the top 10, 20, 30 and 50 most correlated genes to each of the four key transcription factors, enabling the comparison and selection of the optimal number of genes for the downstream programs. As a result, the grid of features was expanded from twenty to eighty groups of genes, following final structure:

- 20 groups consisting of 40 genes each (four downstream programs of 10 genes each)
- 20 groups consisting of 80 genes each (four downstream programs of 20 genes each)
- 20 groups consisting of 120 genes each (four downstream programs of 30 genes each)
- 20 groups consisting of 200 genes each (four downstream programs of 50 genes each)

This approach allowed the exploration and comparison of a range of feature combinations for the downstream programs.

5.3.2 Hyperparameter tuning

Different machine learning algorithms were evaluated to compare performance, including nearest centroid classifier (NCC), k-nearest neighbors (KNN), linear support vector machine (SVM) and random forest (RF). While the classifier performance was estimated by averaging the results of the

20 outer CV runs, the hyperparameter optimization was performed in each of the runs, with the 3-fold inner CV. The best hyperparameter was subsequently selected based on the average f1 performance across the inner 3-fold CV and used to train and evaluate in its corresponding outer subset.

Depending on the specific classification algorithm, the grid of hyperparameters evaluated is presented in Table 3.

Model	Hyperparameter	Levels
NC	N/A	*default*: euclidean distance is used
KNN	k	2, 3, 5, 7, 9, 10, 11
SVM	cost	$9.8e^{-4}$, $5.5e^{-3}$, $3.1e^{-2}$ $1.8e^{-1}$, 1.0, 5.7, $3.2e^{1}$
RF	number trees	300, 600, 900, 1200
	min node size	6, 9, 12, 15
	num. randomly selected predictors	*default*: square of the total num. of predictors

Table 3. Grid of hyperparameters evaluated.

R v4.1.1 was utilized, and the models were implemented using tidymodels framework v1.1.0 and rstatix v0.7.2. The algorithm-specific package requirements consist of lolR v.2.1, kknn v.1.3.1, kernlab v.0.9-32 and randomForest v4.7.1-1 respectively.

5.3.3 Model Selection and Further Evaluation

Models were compared across the different classification algorithms and grid of features. Multiple metrics were used for comparison, including accuracy, f1, precision (positive predictive value), recall (sensitivity), specificity and negative predictive value. For each model and performance metric, the result was estimated by averaging the results across the 20 outer CV runs.
After model comparison, the combination of algorithm and features was selected based on the f1 performance, and the final NAPY classifier for application on further data sets has then been generated by training on the full 80% of nested CV data.

The final NAPY classifier was then evaluated in the remaining 20% of spared-out samples that were never used for training.

As a further validation step in an independent dataset generated from a different type of biospecimen, our NAPY classifier was evaluated in the 48 human SCLC cell lines from CCLE. Refer to Figure 10 for the summary of the complete pipeline.

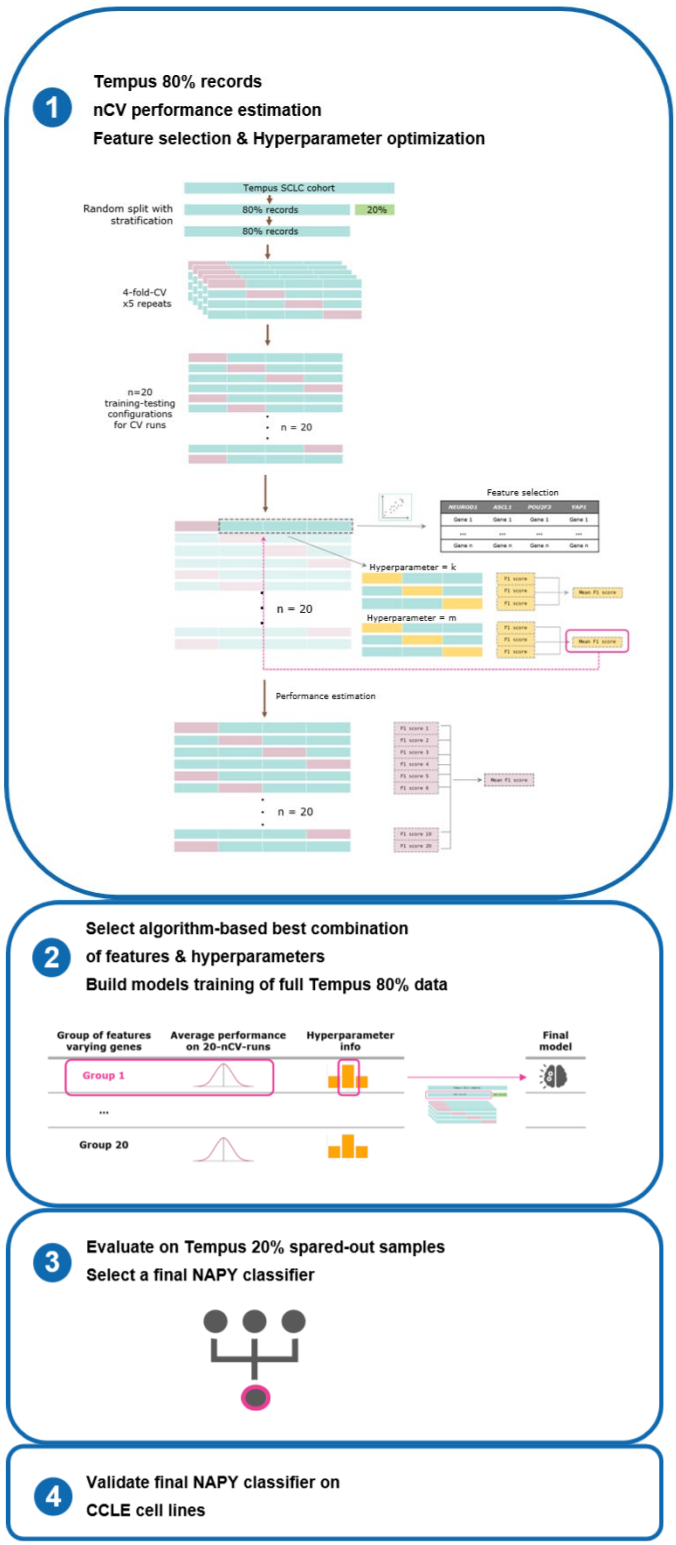

Figure 10. Summary of complete machine learning pipeline: features selection, model evaluation, model selection, prediction of Tempus 20% spared-out data, further validation in independent dataset.

5.4. Additional expression patterns by subtype

A compendium of 320 gene expression signatures were selected for evaluation from which, 313 cancer-related signatures were taken from Kreis *et al.* (2021) (22), 4 immune cell infiltration signatures from Aybey *et al.* (2023) (29), 1 Notch signature from Braune *et al.* (2024) (30), and 2 other interferon signatures developed in Merck not yet published.

In order to evaluate the translatability of these expression signatures developed in other datasets to the Tempus SCLC cohort, the coherence score for each gene signature was computed on the Tempus SCLC cohort. According to the RosettaSX study by Kreis *et al.* (2021), a gene expression signature developed in a different dataset is deemed translatable to a new dataset if its coherence score in the new dataset is higher than 0.2 units (22).

The coherence score for each signature was calculated by averaging the pairwise Pearson correlation coefficients between genes in the Tempus SCLC cohort, considering that at least 80% of the genes within the signature were present in the dataset (22).

The above strategy is supported by the idea that the translatability of a signature on a new dataset can be assessed by analysis of the correlation structure of the signature's genes in the new data set. If there are strong correlations between the genes of a module/signature, it suggests that there are samples within the dataset that exhibit high expression of those genes. This indicates a common positive transcriptional regulation of the module, which may be attributed to various biological factors such as the state of cellular differentiation or the activity of an upstream pathway. Conversely, the absence of correlations could be due to biological irrelevance, technical noise, or differences in patient characteristics within the dataset as proposed by Kreis *et al.* (2021).

After obtaining the expression signatures translatable to our Tempus SCLC cohort, the score was calculated for each of them in each sample. The score was computed by averaging the z-scale expression of the genes within each expression signature, for each sample. Assuming there are 100 translatable signatures and a dataset consisting of 300 samples, a total of 3000 scores would be obtained.

Signatures were then filtered for those showing significant expression differences between subtypes, as determined by an ANOVA with an uncorrected p-value < 0.05 used as a selection heuristic. Wilcoxon tests on the signature scores vs molecular subtypes were then performed in a pairwise manner to assess significant expression variance across subtypes, with p-values adjusted using the Bonferroni method.

5.5. Therapeutic implications by subtype

Clinical information of the patients whose molecular data was used for the development of the machine learning classifier is available in the Tempus dataset. Clinical information includes line of treatments and medications received as well as recurrence and progression of the disease. This information was used to evaluate the variability in treatment selection over time and versus the molecular subtype.

An overall survival (OS) analysis was performed considering the date of primary diagnosis as the start date and evaluating the number of events (deaths) over a fixed period of time. The analysis was performed both for the patients that received chemotherapy in combination with immunotherapy (treatment specific), and independent of the line of treatment (treatment unspecific). The OS was compared across the molecular SCLC subtypes. To avoid biased results related to stage of the disease at the time of diagnosis, only patients with stage 4 of the disease at the time of diagnosis

were considered for the analysis. The analysis was implemented with survival::survfit() function version 3.2-11, which estimates the empirical survival distribution with the Kaplan-Meier method.

The Kaplan-Meier method estimates the probability of a patient's survival beyond a specified time. By considering equally spaced intervals of time within the study period, along with the count of participants alive at the start of each interval and the count of those who deceased, the estimator for the probability of survival at a given time (t+1) is calculated using the following formula:

$$S_{t+1} = S_t * (1 - \frac{D_{t+1}}{N_{t+1}})$$

where S_t refers to the probability of survival at time *t*, D_{t+1} refers to the number of events (e.g., deaths) that happened at time *t+1*, and N_{t+1} refers to the individuals known to have survived up to time *t+1.* (31)
In other words, it is the probability of the event (death) at a specific time, given that the patient has survived up to that time.
As time progresses, the survival function shows how the probability of survival decreases.

In order to determine the influence of the subtype variable in the survival time, the Cox Proportional Hazards Survival Regression (or Crox Regression) statistical method was used. This regression is designed to assess the effect of the variable subtype on the shape of the survival curve.

$$\lambda(t) = \lambda_0(t)^{(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}$$

where $x_k$are the variables and β_k are the coefficients of the variables.

By looking at the estimated regression coefficients and their associated p-value, the influence of the variable in the survival can be determined with its statistical significance. The null hypothesis is that the coefficient is zero, meaning that the variable in study does not have an influence on survival.
Therefore, a positive coefficient indicates an increase in the log hazard of death (higher risk) and a negative coefficient indicates a decrease in the log hazard (lower risk), coefficients around zero indicate no or very little effect on the hazard. The hazard ratio (without the logarithm) can be quantified by taking the exponentiated form of the coefficient and helps to directly interpret the relative change in the probability of death compared to a reference. In our analysis, the survival of subtype A is used as reference in the calculation of the hazard ratio. The analysis was implemented with survival::coxph() function version 3.2-11.

6. Results and discussion

6.1 Subtype Assignment based on master TFs expression

From the 460 Tempus SCLC tumor samples, 332 were assigned to a specific subtype: 105 SCLC-N (32%), 111 SCLC-A (33%), 44 SCLC-P (13%) and 72 SCLC-Y (22%). The proportion of each subtype in the complete dataset was consistent with other reported studies (15). From the 332 labeled tumor samples, 80% were used for training the classifier in a nested cross-validation manner and the remaining 20% of the samples were left aside for final classifier testing.

6.2 Descriptive analysis of clinical features in Tempus SCLC cohort

Patients were almost evenly distributed between females (52%) and males (48%), mostly white ethnicity and ex-smoker or current smoker if information was available. Subjects were mostly from an age range of 60-69. Samples were mostly in stage 4 which typically represents an advanced cancer that has spread to distant organs or tissues. Table 4 contains the available variables and its distribution. Supplementary Table 2 contains oncology definitions, abbreviations and nomenclature related to Table 4.

Characteristic	N = 332[1]	Characteristic	N = 332[1]
Sample disease: recurrence		**Subject sex**	
primary	322 / 322 (100%)	Female	172 / 332 (52%)
Unknown	10	Male	160 / 332 (48%)
Sample disease: grade		**Subject ethnicity**	
Grade 3 (poorly differentiated)	38 / 89 (43%)	White	217 / 247 (88%)
High Grade	35 / 89 (39%)	Black or African American	22 / 247 (8.9%)
Grade 4 (undifferentiated)	12 / 89 (13%)	Other Race	5 / 247 (2.0%)
Grade 2 (moderately differentiated)	3 / 89 (3.4%)	Asian	3 / 247 (1.2%)
Grade 1 (well differentiated)	1 / 89 (1.1%)	Unknown	85
Unknown	243	**Subject age**	
Sample disease stage: Metastasis		60-69	147 / 331 (44%)
M1	81 / 106 (76%)	70-79	91 / 331 (27%)
M0	22 / 106 (21%)	50-59	62 / 331 (19%)
MX	3 / 106 (2.8%)	<=49	18 / 331 (5.4%)
Unknown	226	80-89	13 / 331 (3.9%)
Sample disease stage: Node involvement		Unknown	1
N2	43 / 109 (39%)	**Subject prebiopsy smoking status**	
N3	35 / 109 (32%)	Current smoker/Ex-smoker	109 / 115 (95%)
N0	11 / 109 (10%)	Never smoker	6 / 115 (5.2%)
NX	11 / 109 (10%)	Unknown	217
N1	9 / 109 (8.3%)	**Subject prebiopsy lines of therapy completed**	
Unknown	223	1	39 / 51 (76%)
Sample disease stage: Tumor		>1	12 / 51 (24%)
T4	43 / 108 (40%)	Unknown	281
T3	21 / 108 (19%)	**Subject prebiopsy outcomes last recorded**	
T1	15 / 108 (14%)	Progressive Disease	54 / 68 (79%)
T2	15 / 108 (14%)	Complete Response	7 / 68 (10%)
TX	12 / 108 (11%)	Partial Response	5 / 68 (7.4%)
T0	2 / 108 (1.9%)	Stable Disease	2 / 68 (2.9%)

Unknown	224	Unknown	264
Sample disease stage: Stage			
Stage 4	207 / 237 (87%)		
Stage 3	26 / 237 (11%)		
Stage 2	3 / 237 (1.3%)		
Stage 1	1 / 237 (0.4%)		
Unknown	95		

[1] n / N (%)

Table 4. Descriptive analysis on the 332 SCLC Tempus cohort.

6.3 Machine Learning Results

6.3.1 Grid of downstream programs used as predictors

A grid comprising 80 predictor groups was generated by varying the length of the gene expression signatures, obtained from the correlations measured in the training part of each of the 20 outer cross-validation iterations (see methods). Supplementary Data Table 3 shows the 80 groups of genes evaluated as predictors in the machine learning models.

6.3.2 Comparison of algorithms performances

We built 80 slightly different models for each of the algorithms NCC, KNN, SVM, and RF by training on each of the 80 groups of predictor genes. The 20-outer CV runs enabled us to estimate the classification performance for each of the models.

To evaluate whether a particular algorithm was in general better than others in classifying the SCLC subtypes, we compared the f1 score, sensitivity (recall), specificity, precision (positive predictive value) and negative predictive value distributions derived from these 80 models, across the different algorithms.

Table 5 provides the mean, median, and quantiles of these performance distributions for each algorithm and metric.

metric	mean	q0.25	median	q0.75
k nearest neighbors				
f1 score	0.83	0.82	**0.83**	0.83
neg. pred. value	0.95	0.95	0.95	0.96
precision	0.87	0.87	0.87	0.88
sensitivity	0.81	0.80	0.81	0.81
specificity	0.95	0.95	0.95	0.95
random forest				
f1 score	0.83	0.82	**0.83**	0.83
neg. pred. value	0.96	0.96	0.96	0.96
precision	0.87	0.87	0.87	0.87
sensitivity	0.81	0.81	0.81	0.81
specificity	0.95	0.95	0.95	0.95
support vector machine				
f1 score	0.84	0.83	**0.84**	0.84
neg. pred. value	0.96	0.96	0.96	0.96
precision	0.88	0.88	0.88	0.89
sensitivity	0.82	0.82	0.82	0.83
specificity	0.95	0.95	0.95	0.95
nearest centroid				
f1 score	0.82	0.81	**0.82**	0.82
neg. pred. value	0.95	0.95	0.95	0.95
precision	0.87	0.86	0.87	0.88
sensitivity	0.81	0.80	0.81	0.81
specificity	0.95	0.94	0.95	0.95

Table 5. Performance distributions (25 quantile, mean, median, 75 quantile) for the four algorithms evaluated: nearest centroid, k-nearest neighbors, support vector machine and random forest. Each distribution generated from the results of 80 models trained on the 80% of Tempus patient records, subtly varying the predictor genes according to the generated grid. These results were estimated using 20-outer CV runs.

Considering the f1 score performance metric, SVM exhibits a higher median value (84%) compared to the other algorithms (82-83%). However, the variation in medians among the algorithms is minimal, with differences of no more than 2% units. Similarly, for the other performance metrics, the median values across algorithms differ only in 1% units, all surpassing the 80% threshold.

6.3.4 Benchmarking in the best number of features

To assess the impact of varying the number of genes used to train the models, the performance results were compared across the different program's length used. In this case, the f1 score metric was utilized for the comparison. For each algorithm (NCC, KNN, SVM, and RF), the 80 data points from the distribution were separated into four independent distributions of 20 points each, based on the number of features employed to train the model.

The number of features to train the model could be:

- 40 features : 4 subtype-specific downstream programs, of 10 genes each.
- 80 features: 4 subtype-specific downstream programs, of 20 genes each.
- 120 features: 4 subtype-specific downstream programs, of 30 genes each.
- 200 features: 4 subtype-specific downstream programs, of 50 genes each.

For each algorithm, four distributions were subsequently compared. This comparison aimed to determine if there were important differences in the classifier's performance based on the number of features used for training.

Figure 11 shows the comparison of the classifier´s f1 performance score, based on the downstream program´s length used to train the model. The comparison is done within each algorithm independently.

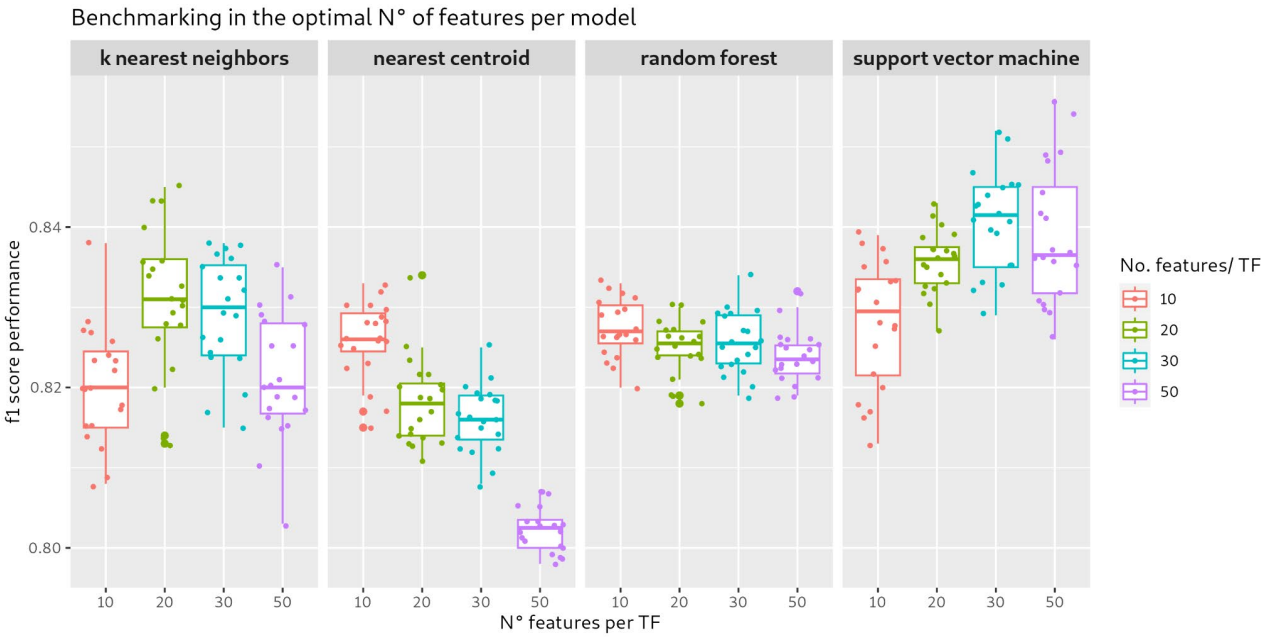

Figure 11. F1 performance score distribution across the different downstream program´s length used to train the model, for each of the algorithms. X-axis represents the number of genes per expression signature. Each model uses 4 TF-associated expression signatures as predictors.

For the k nearest neighbors classifiers, the best performances were obtained using either 80 features (4 downstream programs of 20 genes each) or 120 features (4 downstream programs of 30 genes each).

In the case of nearest centroid classifiers, the best performances were obtained using 40 features (4 downstream programs of 10 genes each). On the contrary, using 200 features (4 downstream programs of 50 genes each) lead to the worst f1 scores by far.

When analyzing random forest classifiers, it was observed that the f1 average score demonstrates low variability regardless of the number of features used to train the model. This outcome might be attributed to the methodology employed by the random forest algorithm, where in each step, the predictor that minimizes the classification error rate is chosen to build the tree. Depending on the complexity and depth of the tree, not all variables may be used in its construction. Consequently, despite we increase the number of predictors, the model's use of predictors may be constrained by the depth of the tree, and we would expect the importance of the predictors not to vary much from model to model.

For the support vector machine classifiers, inferior performances were observed when using 40 features (4 downstream programs of 10 genes each), while differences in f1 performance scores

between SVM models using 80, 120, and 200 features were not substantial. On the other hand, there is a greater intra-group variability in the f1 scores for models trained with 200 features and 120 features compared to those trained with 80 features, suggesting higher stability for the latter.

6.3.5 Hyperparameter tuning

For each of the 20-outer cross-validation runs, a grid of hyperparameters was evaluated using a 3-fold inner cross-validation (see methods). The hyperparameter that resulted in the highest average F1 score across the inner 3-fold cross-validation was selected as the best hyperparameter and used for the outer training.

In order to examine how the accuracy of the F1 score relates to the hyperparameters, we plotted the average F1 scores obtained from the inner 3-fold cross-validation for each hyperparameter. This analysis was conducted for a fixed set of features and algorithm.

In the plots, the gray lines represent the relationship obtained from each of the 20 outer iterations. The blue smooth line represents the locally weighted polynomial regression (loess) fitted to the F1 score from those 20 outer iterations. The gray shadow around the smooth blue line represents the confidence interval. Finally, the best hyperparameter resulting from each of the 20 outer iterations is highlighted with a dot.

Figure 12 shows the relationship between F1 score performance and different values of k for the k-nearest neighbor classifier. The greatest performance increase is observed up until k=9.

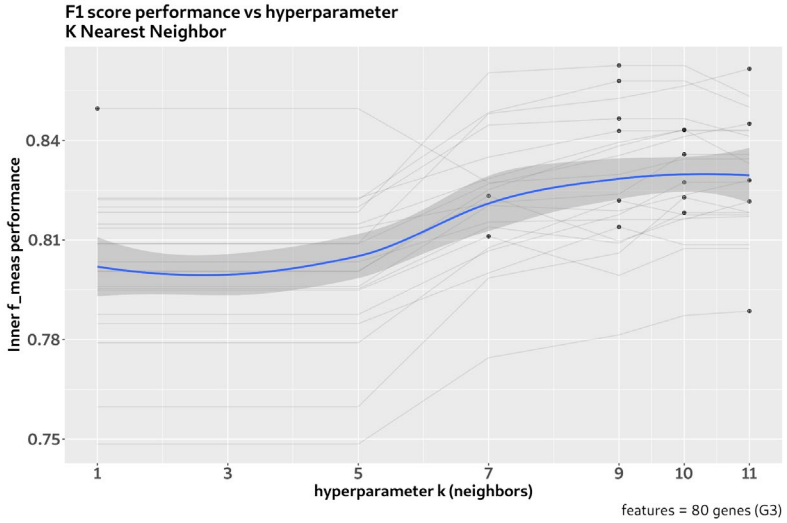

Figure 12. F1 performance score across the different k hyperparameters evaluated in the inner loop cross validation for the KNN classifier trained with 80 features (group 3).

Figure 13 shows the relationship between F1 score performance and different values of the cost margin for the support vector machine classifier. The greatest performance is observed with c=0.031.

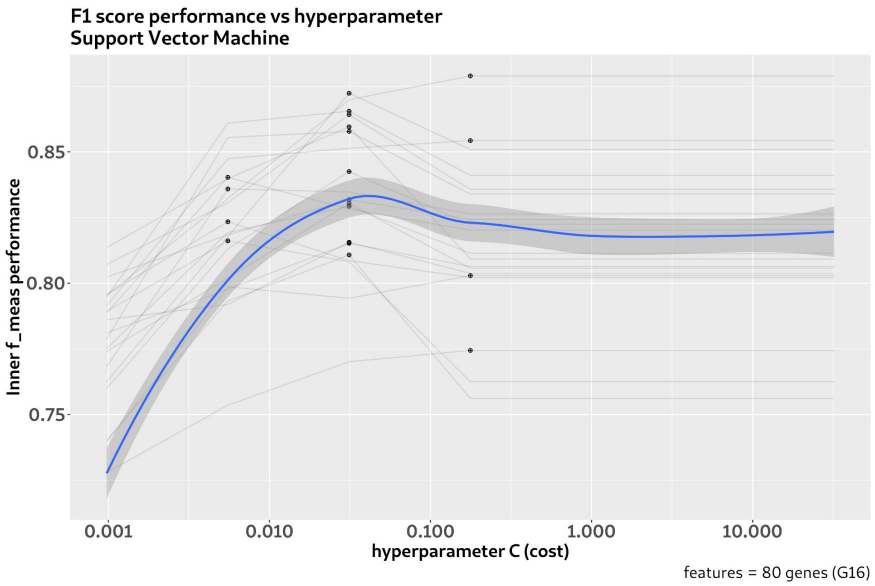

Figure 13. F1 performance score across the different c hyperparameters evaluated in the inner loop cross validation for the SVM classifier trained with 120 features (group 12).

Figure 14 shows the relationship between F1 score performance and different combinations of number of trees and minimal node size for the random forest classifier. In this case, the number of trees is represented in the x axis and the different minimal node sizes are represented with different colors. The variability in F1 score performance is not high among the different combinations of hyperparameters.

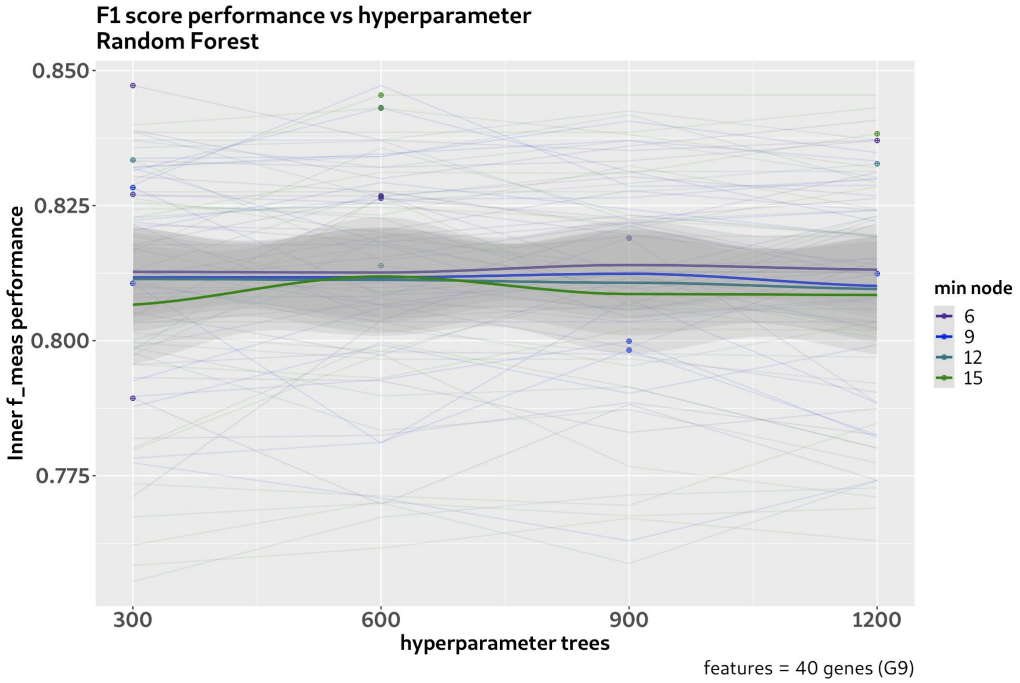

Figure 14. F1 performance score obtained in the inner loop cross validation across the different values for the number of trees hyperparameter for the RF classifier trained with 40 features (group 9). The correlation with the different values for minimal node size hyperparameters represented with colors.

6.3.6 Model comparison

From the models built from each algorithm (NC, KNN, SVM and RF), a combination of features and hyperparameters (if applicable) was chosen to train a final model on the full 80% of Tempus nested CV data. These four final NAPY classifiers where then evaluated in the remaining 20% of Tempus spared-out data, which was never been used for training.

The features for each final model were selected based on the algorithm-specific optimal number of features. If there is no number of features that yields significantly better results, the simplest model is chosen, meaning the one with the lowest number of features. Among the groups with the same number of genes, one was selected from those that achieved the best F1 performance score in the cross-validation training.

For the NC algorithm, the predictors were selected based on the highest f1 performance score among all the groups of predictors with 40 genes (4 downstream programs of 10 genes each). This decision was made because nearest centroid classifiers exhibited higher performance when using 40 predictors.

In the case of KNN algorithm, the predictors were selected based on the highest f1 performance score among all the groups of predictors with 80 genes (4 downstream programs of 20 genes each). On the other hand, the k hyperparameter was defined as the most frequent value that was selected as best hyperparameter in the 20 cross-validation iterations. Refer to Figure 15 for the hyperparameters frequency analysis.

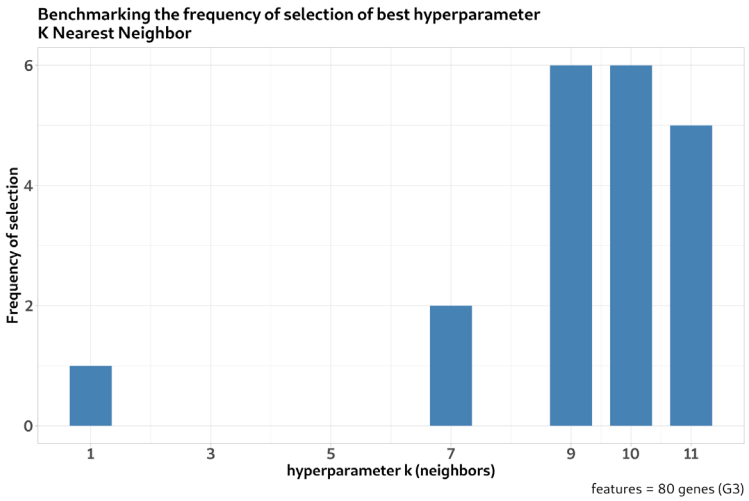

Figure 15. Frequency of selection of k as the best hyperparameter in the 20 runs of the outer loop for the KNN classifier trained on 80 features (group 3).

For the SVM algorithm, a group of predictors was selected among all the groups of predictors with 80 genes (4 downstream programs of 20 genes each). As in the case of the KNN algorithm, the cost hyperparameter was defined as the most frequent value that was selected as best hyperparameter in the 20 cross-validation iterations. Refer to Figure 16 for the hyperparameters frequency analysis.

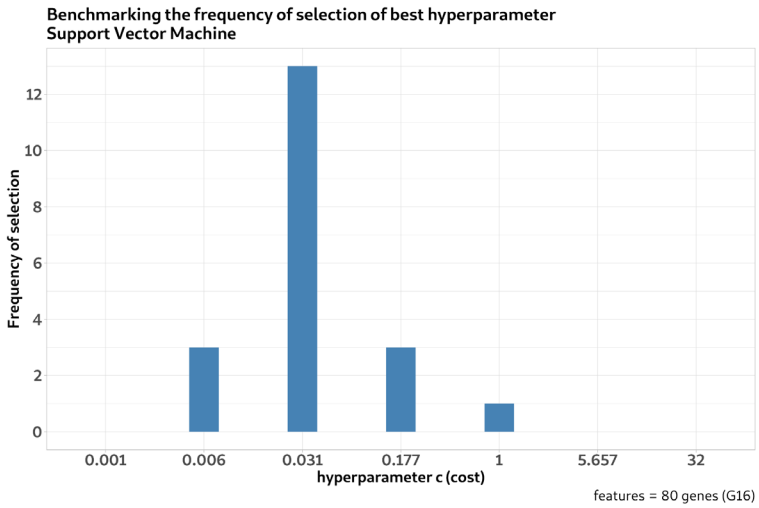

Figure 16. Frequency of selection of c as the best hyperparameter in the 20 runs of the outer loop for the SVM classifier trained on 120 features (group 12).

Regarding RF algorithm, the predictors were selected based on the highest f1 performance score among all groups containing 40 genes (4 downstream programs of 10 genes each). This selection was made since the other downstream programs' length did not yield significantly better performance. The number of trees and minimal node size hyperparameters combination was chosen based on both, frequency at which the combination appeared as the best hyperparameters and the aim for trees with lower complexity. Refer to Figure 17 for the hyperparameters frequency analysis.

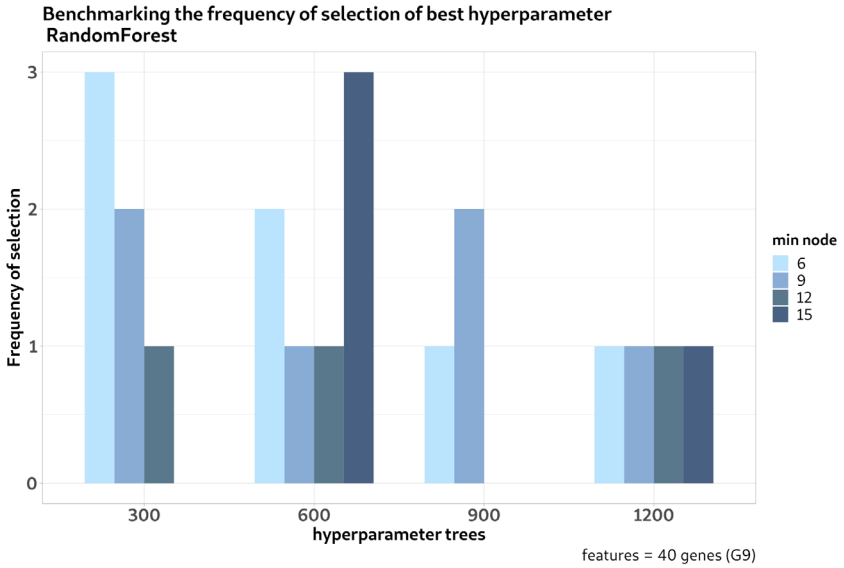

Figure 17. Frequency of selection as the best hyperparameter combination (number of trees & minimal node size) in the 20 runs of the outer loop for the RM classifier trained on 40 features (group 9).

Table 6 compares the performance results of the four final NAPY classifiers obtained from the 20-outer CV training runs on the 80% of Tempus cohort as well as the results obtain from the evaluation in the remaining 20% of Tempus spared-out samples that were never used for training.

metric	nearest centroid		k nearest neighbor		support vector machine		random forest	
	training 80% 20-outer cv	testing 20% spared-out	training 80% 20-outer cv	testing 20% spared-out	training 80% 20-outer cv	testing 20% spared-out	training 80% 20-outer cv	testing 20% spared-out
accuracy	0.86	0.90	0.88	0.87	0.87	0.90	0.87	0.90
f1 score	0.83	0.88	0.85	0.86	0.84	0.89	0.83	0.88
neg pred value	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.97
pos pred value	0.89	0.91	0.90	0.90	0.88	0.92	0.87	0.92
precision	0.89	0.91	0.90	0.90	0.88	0.92	0.87	0.92
recall	0.82	0.87	0.83	0.84	0.82	0.87	0.82	0.86
sensitivity	0.82	0.87	0.83	0.84	0.82	0.87	0.82	0.86
specificity	0.95	0.96	0.96	0.95	0.95	0.96	0.95	0.96

NC: 40 features (G9)
KNN: 80 features (G3), hyperparameter: k = 9
SVM: 80 features (G16), hyperparameter: cost = 0.031
RF: 40 features (G9), hyperparameters: trees = 300, min. node = 6

Table 6. Classification performance results for the four final NAPY classifiers. Nearest Centroid model trained on 40 genes from group 9. K-Nearest Neighbor model trained on 80 genes from group 3, hyperparameter k=9. Support Vector Machine model trained on 80 genes from group 16, hyperparameter c=0,031. Random Forest model trained on 40 genes from group 9, hyperparameter trees=300 and min. node=6. For each model, the results estimated from the 20-outer CV training runs on the 80% of Tempus cohort are in the first column; the results from the evaluation in the remaining 20% of Tempus spared-out samples are in the second column.

Looking at the performance of the four final NAPY classifiers in the 20% of Tempus left-out data that were never used for training, all f1 scores were 86% or above, recall/sensitivity 84% or above, precision/positive predicted value 90% or above, specificity reached 95% or above and negative predicted value 96% or above. Performances are similar across models and also between training and testing data.

6.3.7 Model selection and further evaluation in an independent dataset from Cancer Cell Line Encyclopedia (CCLE)

After model comparison, we finally selected the support vector machine (SVM) model built on 80 predictor genes. This choice was based on two factors. Firstly, the median f1 performance scores for the SVM models were consistently among the highest when compared to other algorithms (Table 5, Figure 7). Secondly, the differences in median f1 performance scores between SVM models using 80, 120, and 200 features were not substantial, leading to the selection of a simpler model with 80 features (Figure 7). The specific set of 80 features was chosen among those that yielded the best f1 performance score in the Tempus data. Refer to Table 6 with the final NAPY SVM classifier selected, and its performance on the Tempus data.

In order to have an additional layer of validation of our machine learning model along with its subtype-specific downstream programs, our NAPY classifier was further evaluation in the 48 SCLC cell lines from the CCLE dataset.

Table 7 compares the performance results of our final NAPY SVM classifier obtained from the 20-outer CV training runs on the 80% of Tempus cohort, the results obtain from the evaluation in the remaining 20% of Tempus spared-out samples that were never used for training, and the performance results obtained from the evaluation in the CCLE dataset.

metric	Tempus *80% full training data from nCV* (n=264)	Tempus *20% spared out for final assessment* (n=68)	CCLE *further validation* (n=48)
accuracy	0.87	0.90	0.90
f1 score	0.84	0.89	0.83
neg. pred. value	0.96	0.97	0.96
pos. pred. value	0.88	0.92	0.82
precision	0.88	0.92	0.82
recall	0.82	0.87	0.86
sensitivity	0.82	0.87	0.86
specificity	0.95	0.96	0.97
Final NAPY SVM Classifier, cost=0.031, features=80 genes			

Table 7. Classification performance results obtained with the final NAPY SVM classifier trained on the full 80% of Tempus nested CV records, using the 80 genes from group 16 as predictors and hyperparameter cost=0,031. Tempus training cv: results obtained from the 20-outer CV training runs on the 80% of Tempus cohort. Tempus testing: results obtain from the evaluation in the remaining 20% of Tempus spared-out samples that were never used for training. CCLE validation: results obtained from the evaluation in the CCLE dataset.

The classification results demonstrate the predictive power of the four 20-gene transcriptional downstream programs linked to the key transcription factors. These programs yield SVM models with an accuracy of 90% and a specificity exceeding 95% in the validation dataset, showcasing strong predictive capabilities in independent data. Comparable performance during the nested-CV training phase and the assessment of the final NAPY classifier on left-out data provide evidence for little influence of bias during classifier training.

The per-class performance results and corresponding confusion matrix for the reserved 20% of Tempus records and for the CCLE dataset are outlined in Table 8. The classifier achieved per-class accuracies ranging from 92% to 97% and per-class specificity results ranging from 91% to 100%, measured as one-vs-rest binary classification. Specifically, of the 68 samples from the reserved 20% of Tempus records, 61 were correctly classified, whereas 43 out of 48 samples from CCLE dataset were correctly classified.

A.

Tempus 20% records save for assessment *Per-class performance metric*				
metric*	**A**	**N**	**P**	**Y**
accuracy	0.93	0.96	0.97	0.94
sensitivity	0.96	0.95	0.78	0.80
specificity	0.91	0.96	1.00	0.98
precision	0.85	0.91	1.00	0.92
recall	0.96	0.95	0.78	0.80
f1 score	0.90	0.93	0.88	0.86
neg. pred. value	0.98	0.98	0.97	0.95
pos. pred. value	0.85	0.91	1.00	0.92
* Binary estimators Final NAPY SVM Classifier, cost=0.031, features=80 genes				

B.

		Truth			
		A	N	P	Y
Predicted	A	**22**	**1**	**0**	**3**
	N	**1**	**20**	**1**	**0**
	P	**0**	**0**	**7**	**0**
	Y	**0**	**0**	**1**	**12**

C.

CCLE cell lines *Per-class performance metric*				
metric*	**A**	**N**	**P**	**Y**
accuracy	0.96	0.96	0.92	0.96
sensitivity	0.92	0.91	0.75	0.86
specificity	1.00	0.97	0.93	0.98
precision	1.00	0.91	0.50	0.86
recall	0.92	0.91	0.75	0.86
f1 score	0.96	0.91	0.60	0.86
neg. pred. value	0.92	0.97	0.98	0.98
pos. pred. value	1.00	0.91	0.50	0.86
* Binary estimators Final NAPY SVM Classifier, cost=0.031, features=80 genes				

C.

		Truth			
		A	N	P	Y
Predicted	A	**24**	**0**	**0**	**0**
	N	**1**	**10**	**0**	**0**
	P	**1**	**1**	**3**	**1**
	Y	**0**	**0**	**1**	**6**

Table 8. Per-class classification performance results. A-B: Metrics and corresponding confusion matrix for the Tempus 20% left-out set. C-D: Metrics and corresponding confusion matrix for the CCLE set.

The precision/positive predictive value score (TP/(TP+FP)) for the SCLC-P subtype exhibited a low value in the CCLE dataset, possibly due to the imbalanced class distribution (26:11:4:7) and relatively smaller sample size of the SCLC-P class (4 samples) compared to the other classes. We expect that a more balanced distribution with an increased number of SCLC-P samples in the validation set could provide a more precise and more favorable precision/ppv performance score for SCLC-P.

The fact that performance scores for the application of our NAPY SVM models on Tempus cancer patients and CCLE cell lines are predominantly between 80-100% is a remarkable confirmation that our ML-based approach including the four NAPY gene expression signatures as features can be translated between different data sets generated on different types of biospecimen. The four downstream programs linked to the TFs used as predictors in the SVM classifier are shown in Supplementary Data Table 3. Several of these genes have been previously described in other SCLC research studies.

To provide additional insights into the aforementioned results on CCLE data, per-sample signature scores for these CCLE samples were calculated and compared across subtypes. Figure 19 aimed to showcase the variability within subtypes as well as across subtypes for our four downstream programs.

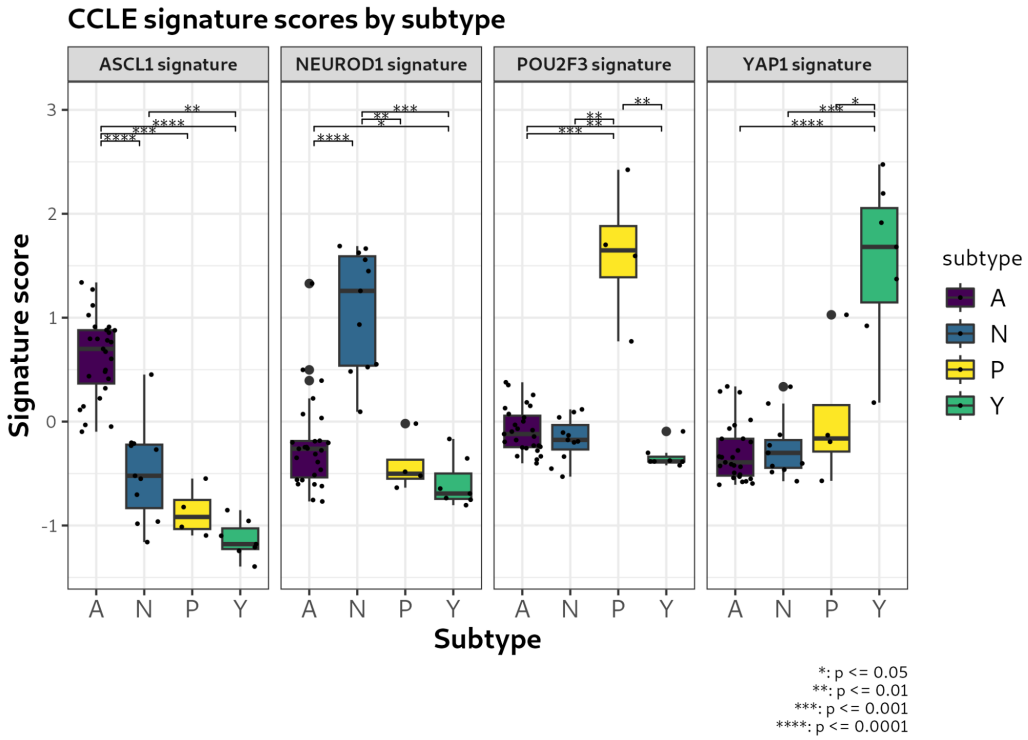

Figure 19. Boxplot of per-sample signature scores versus subtype for the four subtype-specific signatures, assessed on the CCLE cell lines. Subtype pairwise variability evaluated using the Wilcox test, with FDR method applied to adjust the p-values.

Consistent with the classification results, the Wilcoxon comparison of signature scores vs subtypes further confirms that the subtype-specific signature is significantly higher expressed in its corresponding subtype. Therefore, we can expect the SCLC-A samples to have higher scores in the *ASCL1* signature compared to the other SCLC subtypes.

6.3.8 NAPY Classifier subtype-specific gene expression signatures

The 4 downstream programs of 20 genes each, associated with each of the key transcription factors used as predictors in the final support vector machine classifier are shown in Table 9.

Several of these genes have been previously described in other research studies that aim to characterize the transcriptional profiles of SCLC.

	ASCL1	NEUROD1	POU2F3	YAP1
1	SEC11C	CERKL	C11orf53	LATS2
2	DDC	SSTR2	GFI1B	WWTR1
3	RIMKLA	CHRNB4	TRPM5	LAMB2
4	CNKSR3	CHRNA3	APOBEC1	OSMR
5	SCN3A	NHLH1	FAM150A	CYBRD1
6	PTPRN2	NEUROD4	IL19	GPX8
7	CACNA1A	LMO1	PTPN18	MSRB3
8	RGS17	PPP1R17	HES2	ITGB5
9	STK32A	NEUROD2	CALML5	SYDE1
10	GRP	CNTN2	IMP4	RBMS3
11	CA8	SLC17A6	BMX	PLA2R1
12	NOL4	FNDC5	ART3	WTIP
13	RAB3B	SHF	LANCL3	EPHA2
14	SMPD3	THSD7B	PVRL4	MYL9
15	FBLN7	DACH1	KLHDC7A	PMP22
16	MS4A8	HPCA	CALHM3	ZCCHC24
17	ETS2	PROKR1	KIAA1024L	EHD2
18	DLL3	GNG8	ADAMTS19	IFITM3
19	SLC36A4	KIAA1614	ASCL2	AXL
20	NR0B2	CLVS1	COLCA2	THBS1

Table 9 Subtype-specific gene signatures for the final selected model.

ASCL1 expression has been reported to be associated with that of *SEC11C* (5), *DDC* (32) (33) (34), *RIMKLA* (5), *SCN3A* (5), *PTPRN2* (5), *CACNA1A* (33), *RGS17* (5), *GRP* (16) (34) (33) (5), *CA8* (5), *NOL4* (5), *SMPD3* (5), *FBLN7* (5), *MS4A8* (35) (34) (18), *ETS2* (5), *DLL3* (2) (15) (36), *SLC36A4* (5), *NR0B2* (18).

The expression of ***NEUROD1*** is reported to show a positive correlation with that of *CERKL* (18), *SSTR2* (15) (33), *CHRNB4* (33), *CHRNA3* (16) (33), *NHLH1* (10) (33), *NEUROD4* (10) (32) (33), *LMO1* (5), *PPP1R17* (18), *NEUROD2* (18) (33), *CNTN2* (16), *SLC17A6* (5), *FNDC5* (5), *SHF* (5), *THSD7B* (5), *DACH1* (5), *HPCA* (5) (33), *PROKR1* (5), *GNG8* (5), *KIAA1614* (5), *CLVS1* (5).

POU2F3 expression has been found to be linked to that of *C11or53* (4), *GFI1B* (2) (37) , *TRPM5* (17), *APOBEC1* (5), *IL19* (5), *PTPN18* (5), *CALML5* (34), *IMP4* (5), *BMX* (18), *ART3* (5), *LANCL3* (5), *KLHDC7A* (5), *CALHM3* (5), *KIAA1024L(MINAR2)* (38), *ADAMTS19* (5), *ASCL2*(37) (9), *COLCA2* (5).

YAP1 expression has been reported to exhibit a positive correlation with LATS2 (39) (40), *WWTR1* (5), *LAMB2* (5), *OSMR* (32) (40) (41), *CYBRD1* (5), *GPX8* (5) (18) (41), *MSRB* (5) (18), *ITGB5* (5), *SYDE1* (5), *RBMS3* (40) (41), *PLA2R1* (5), *WTIP* (5), *EPHA2* (2) (9) (40) (41), *MYL* (5) (18) (40), *PMP22* (5), *ZCCHC24* (5), *EHD2* (5), *IFITM3* (9) (18) (38) (41), *AXL* (2) (15) (40) (41).

6.3.9 Gene Set Enrichment analysis on the NAPY classifier gene signatures

The four subtype-specific downstream programs underwent a Gene Ontology enrichment analysis using the enrichGO() function from the clusterProfiler package. The objective of this enrichment analysis was to assess if any biological processes or pathways were overrepresented in the gene signature based on the associated genes. A p-value cutoff of 0.1, along with a q-value cutoff of 0.05 using the Benjamini-Hochberg method (also known as the False Discovery Rate), was applied for analysis.

Figure 20 shows the Gene Ontology enrichment analysis result after manually curation to remove redundant and non-specific biological processes.

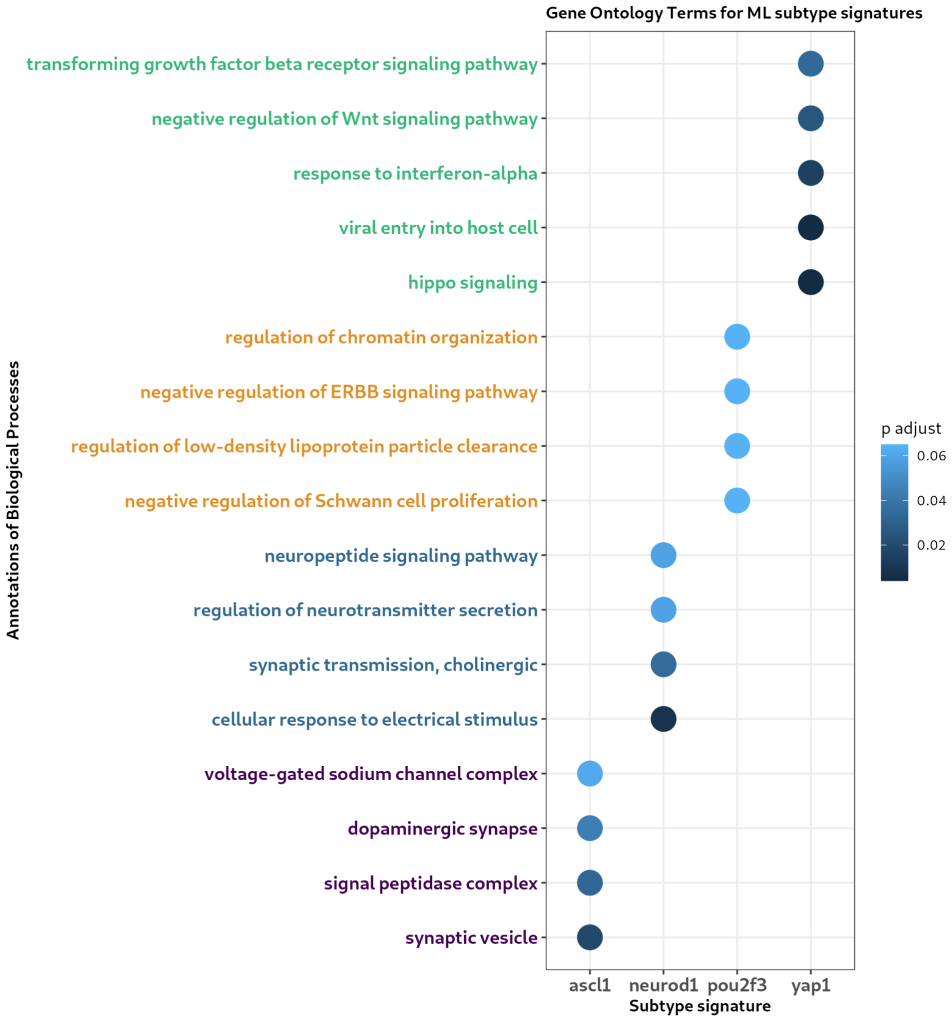

Figure 20. Gene Ontology enrichment analysis on the subtype-specific signatures genes.

Due to their characteristic neuroendocrine phenotype, the *ASCL1*- and *NEUROD1*-associated gene signatures demonstrated enrichment in signaling and synaptic terms. Interestingly, dopaminergic synapses were associated with *ASCL1*, while cholinergic synapses were linked to *NEUROD1*. On the other hand, *POU2F3*-associated signature exhibited enrichment in chromatin organization and negative regulation for *EGFR*, Schwann cell proliferation and low-density lipoprotein clearance. Schwann cells are the main glial cells of the peripheral nervous system and when associated with tumors, *Cao et al.* has demonstrated these cells potentially facilitate the invasion and progression of SCLC (11). In the case of *YAP1*-associated gene signature, ontology terms related to interferon-alfa, TGFB, Hippo and negative regulation of WNT pathway were significantly enriched.

Despite all ontology terms having an adjusted p value above or equal to 0.065, the *POU2F3*-associated gene signature exhibited the least significant gene enrichment among the four.

6.4 Gene expression signatures for cancer pathways are differentially expressed in NAPY subtypes

A compendium of 320 gene expression signatures were selected for evaluation from which, 313 cancer-related signatures were taken from Kreis *et al.* (2021) (22), 4 immune cell infiltration signatures from Aybey *et al.* (2023) (29), 1 Notch signature from Braune *et al.* (2024) (30), and 2 other interferon signatures developed in Merck not yet published.

From the total of 320 gene signatures, 41 were excluded based on either the unavailability of the expression data for at least 80% of the genes within those signatures in our dataset or a coherence score below 0.2, leaving 278 gene signatures for further analysis.

Among these signatures, a total of 258 exhibited significant variance in signature scores across subtypes, as determined through an heuristic analysis of variance ANOVA, with an uncorrected p-value below 0.05.

Subsequently, pairwise comparisons of signature scores between subtypes were performed using the Wilcoxon rank-sum test and applying Bonferroni p-value correction for multiple testing. Only signatures that showed significant differences in at least 3 of the 6 pairwise comparisons, with an adjusted p-value lower than 0.005, were retained. When multiple expression signatures represented the same biological process, in the same context, a representative signature was selected based on the highest F value obtained in the ANOVA analysis. On the other hand, when multiple signatures informed about the same pathway up- or down- regulated, the positive regulated pathways were selected to facilitate result interpretation.

Hence, 128 gene signatures exhibited significant differences in a minimum of three out of six subtype pairwise comparisons of signature scores.

Figure 21 shows a heatmap of the 128 signatures scores by subtype.

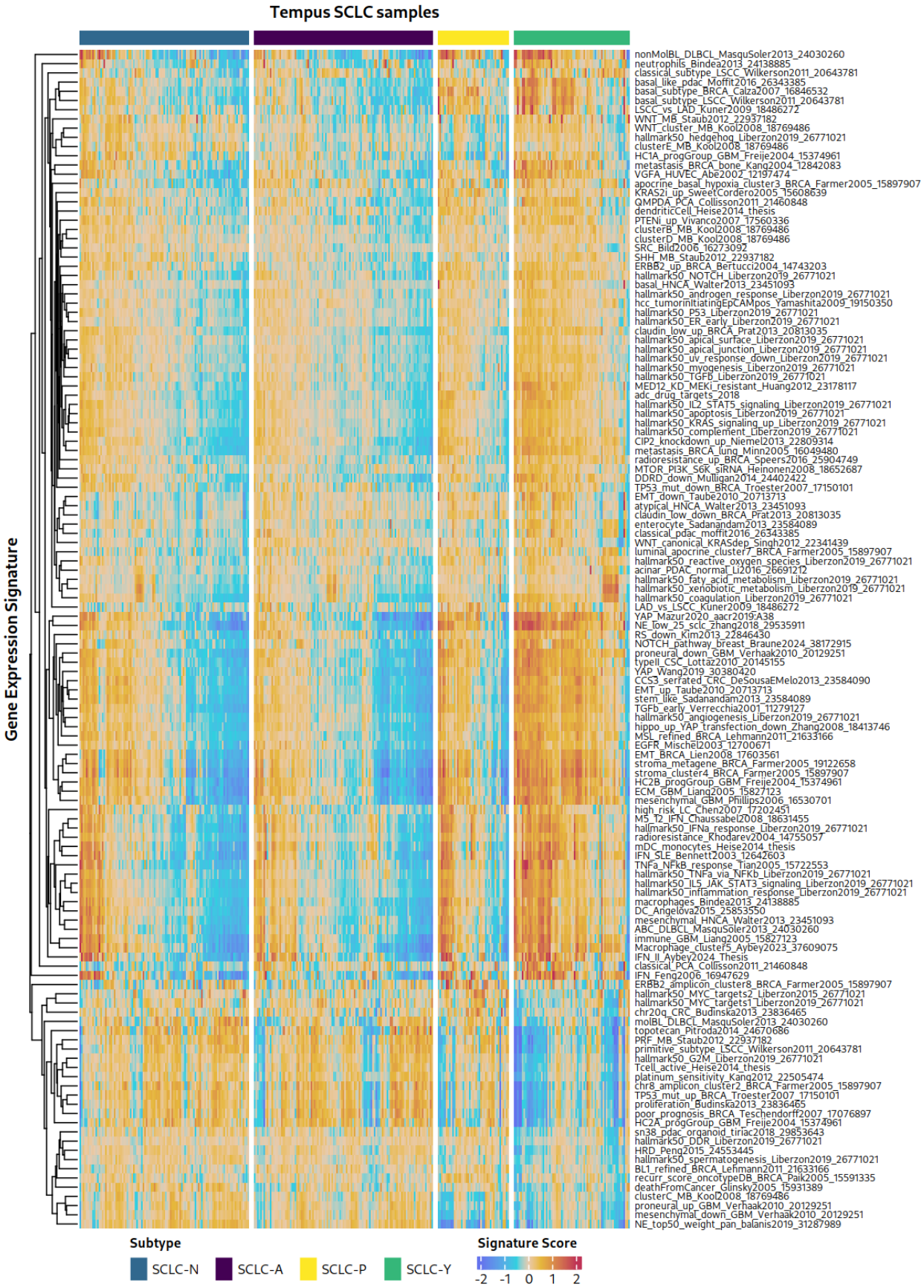

Figure 21. Heatmap of the 128 gene signatures evaluated across the 332 Tempus samples. Columns represent samples, rows represent gene signatures. Orange indicates higher signature scores; blue indicates lower signature scores.

Out of these 120 expression signatures, 35 were selected and are shown in Figure 22 to facilitate the visualization of the patterns across subtypes. These 35 signatures represent biological processes or pathways that have been already described in other SCLC studies, or that are less described in SCLC but may have some clinical or biological relevance.

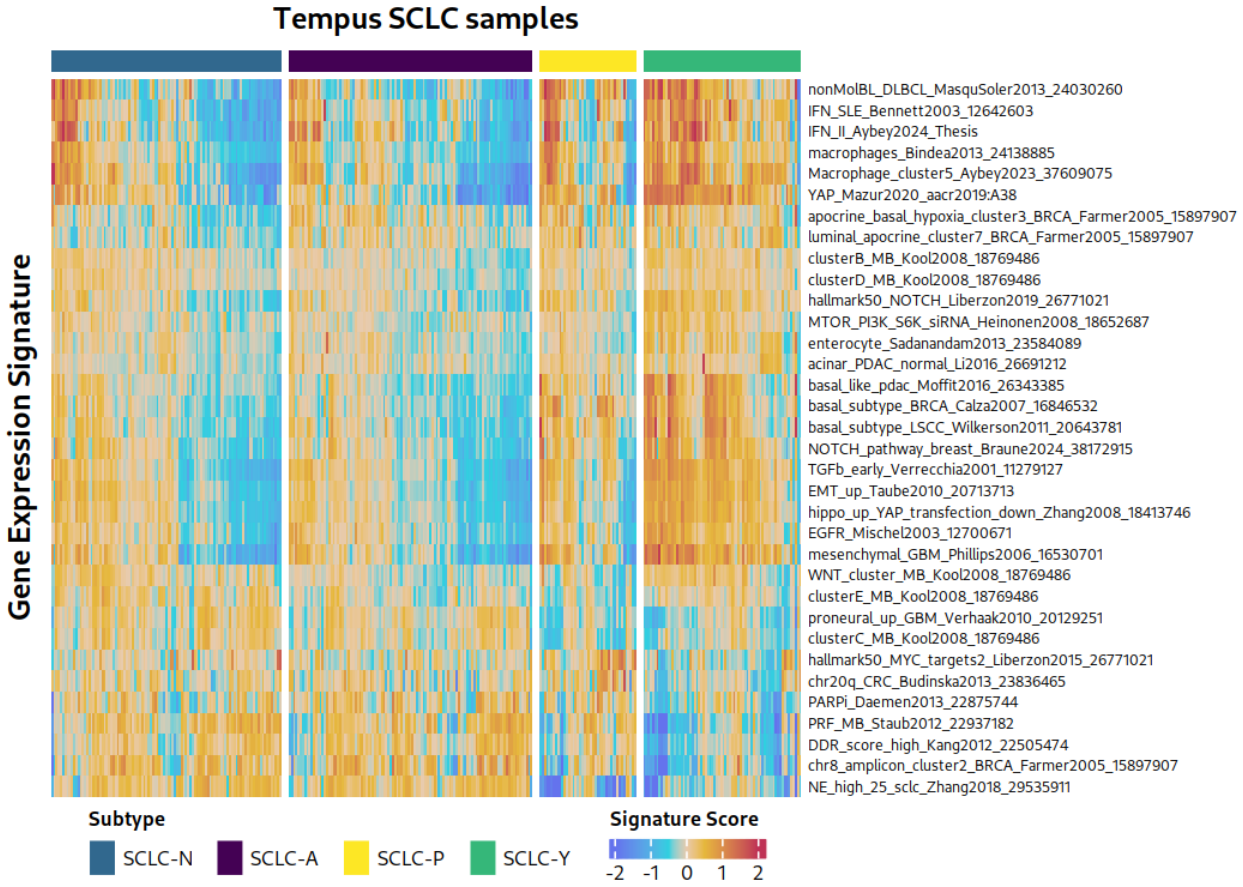

Figure 22. Heatmap of the 35 gene signatures evaluated across the 332 Tempus samples. Columns represent samples, rows represent gene signatures. Orange indicates higher signature scores; blue indicates lower signature scores.

In accordance with previous studies, our findings demonstrate that cancers of the SCLC-N and SCLC-A subtypes exhibit a stronger neuroendocrine gene expression footprint compared to cancers of the SCLC-P and SCLC-Y subtypes, as detected here by significantly higher scores in typical signatures to indicate neural or neuroendocrine phenotypes as detected by the neuroendocrine signature of Zhang et al. and the proneural signature for glioblastoma of Verhaak et al. We observed a similar association for signatures that typically indicate cell cycle activity or cell proliferation in tissues, like the medulloblastoma proliferation signature of Staub et al. or the DNA damage response pathway of Kang et al. A higher cell proliferation activity has been known as feature of neuroendocrine cancer types (9) (15) (16).

The SCLC-Y subtype demonstrates an immune-inflammatory profile, evidenced by significantly higher scores in the interferon signature of Bennett *et al.*, corroborating previous findings, where even the renaming of this subtype from SCLC-Y to SCLC-I has been proposed (15) (16). In addition, macrophage infiltration as measured by our recently published macrophage cell type signature in Aybey *et al.* and by Bindea *et al.* signature, exhibited higher scores for the non-neuroendocrine phenotypes (SCLC-Y and SCLC.P).

Furthermore, our results demonstrate significantly higher scores for the SCLC-Y subtype in several signatures associated with distinct signaling pathway activities. The SCLCs of the SCLC-Y subtype exhibit significantly higher scores in the epidermal growth factor receptor (EGFR) pathway signature of Mischel *et al.*, hippo pathway signature of Zhang *et al*. and PI3K signature of Heinonen *et al.* which is in congruence with several previous reports (9) (40) (43). The activation of the PI3K–AKT–mTOR pathway has been implicated in proliferation and resistance to apoptosis in SCLC (1). The EGFR pathway is a complex network of cell signaling pathways involved in regulating various cellular processes, such as cell growth, proliferation, differentiation, and survival. The SCLC-Y subtype also shows the highest score among all subtypes for YAP signature of Mazur *et al.,* that has been the results of YAP gene dependency in pooled CRISPR screen data, thereby providing independent functional credibility to our SCLC-Y classification.

We observe a similar association with the SCLC-Y subtype for a group of signatures related to a mesenchymal character of tumor tissues: among these are signatures for transforming growth factor-**β** (TGFB) of Verrecchia *et al.*, and epithelial-mesenchymal transition (EMT) of Taube *et al.* While SCLC-Y tumors display the highest scores in these signatures, also the SCLC-P subtype cancers often yield high signals. For the EMT signature, we find that SCLC-Y has the strongest mesenchymal expression characteristics. SCLC-A/-N tumors are rather epithelial-like which is in concordance with other studies (9) (15) (16). Given that TGFB is known to induce epithelial-mesenchymal transition (EMT), it is expected to observe analogous patterns for these two signatures.

We found that a WNT pathway signature (Kool et al., 2008) scores significantly higher for SCLC-N and SCLC-Y subtypes compared to the other subtypes. Alterations leading to the upregulation of the WNT pathway have been frequently reported in relapsing SCLC and associated with chemoresistance acquisition (1). This chemoresistance has been reported to be associated with the non-neuroendocrine phenotypes and mesenchymal SCLC variants (15).

On the other hand, the neuroendocrine phenotypes (SCLC-A and SCLC-N) exhibited lower scores for the Notch signaling pathway signature of Braune *et al.* These results are consistent with the fact that Notch signaling is a key negative regulator of neuroendocrine differentiation in SCLC. Several Notch receptors are negatively correlated with the neuroendocrine score, as is *HES1*, a negative regulator of NE differentiation (9).

Myc family members (*MYC*, *MYCL*, and *MYCN*) are frequently amplified or over-expressed in SCLC tumors and cell lines, with MYC expression strongly negatively correlated with the neuroendocrine phenotypes (9). Our findings indicate that the MYC hallmark signature of Liberzon *et al.* is predominantly highly expressed in the SCLC-P subtype (15) and correlates well with proliferation signatures. These SCLC-P tumors are known to develop from a non-neuroendocrine chemosensory cell type called Tuft cells (34). Some studies have already proposed a special role for the MYC pathway in SCLC-P tumors which is supported by the results for our SCLC cohort (10) (34) (44).

The detailed comparison of the aforementioned expression signatures by subtype is shown in Figure 23.

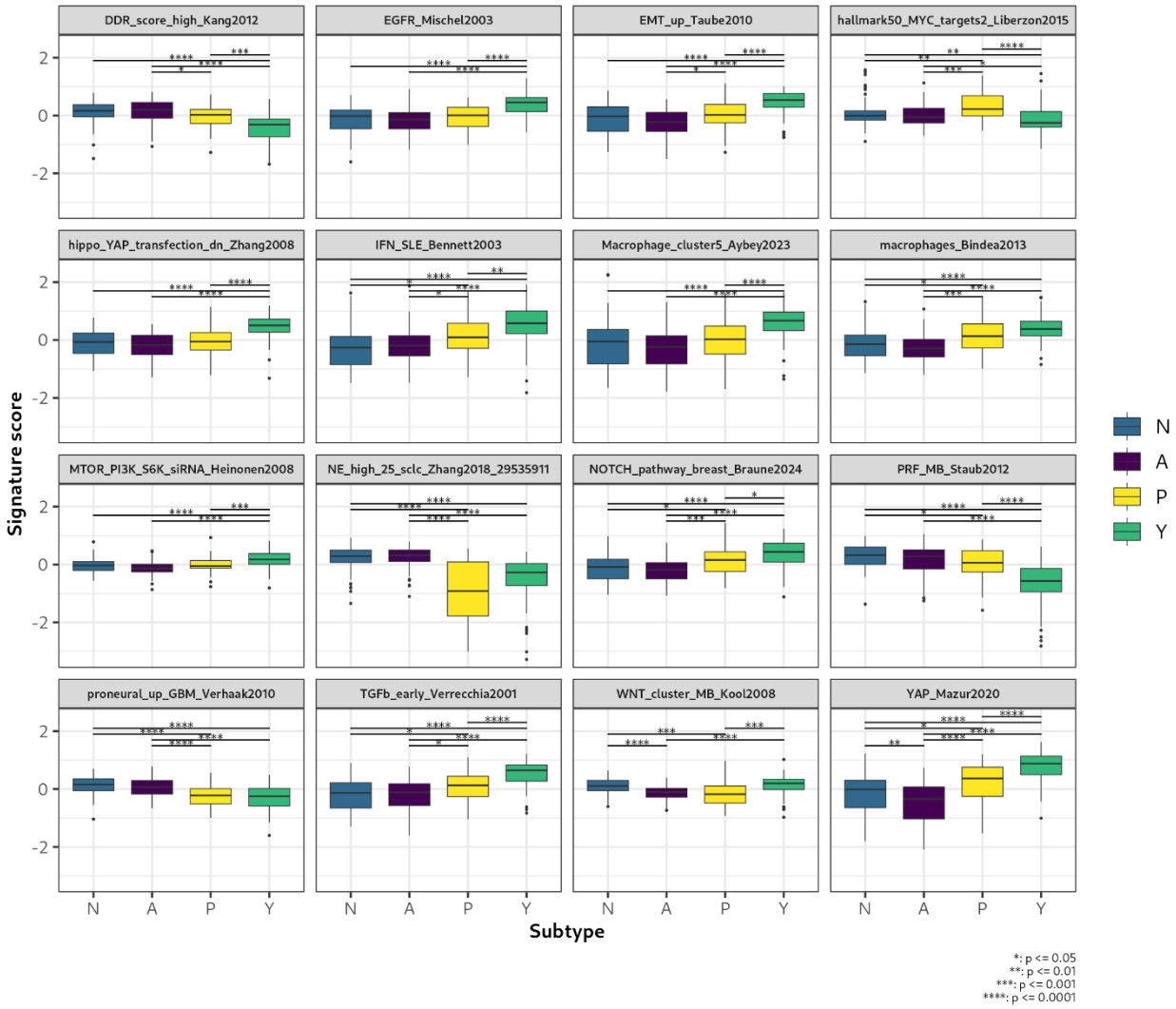

Figure 23. Boxplot of the signature scores across subtypes, for each of the 16 gene signatures. Subtype pairwise variability evaluated using the Wilcox test, with Bonferroni method applied to adjust the p-values.

6.5 Most genomic alterations are non-mutually exclusive across subtypes

Regarding the genomic alterations observed in these 332 SCLC tumor samples obtained from Tempus (Figure 24), the mutational profiles align with findings from previous studies, exhibiting heterogeneous mutational profiles and nearly universal inactivation of *TP53* (95%) and *RB1*(74%) genes (45). In addition, *LRP1B* was also identified with relative high frequency mutation rates (27%). *LRP1B* is presumed to be a tumor suppressor gene and its mutation have already been reported in SCLC and other cancers and has been proposed to have a significance in SCLC progression (46) (47).

We found mutations in KMT2D and KMT2C in all SCLC subtypes at an overall frequency of 17% and 15%, respectively. Both genes are associated with chromatin remodeling pathways and its alterations described in SCLC to contribute to inactivation of tumor suppressors (1) (36) (47).

Like other studies, we observed alterations in *PTEN* (14%) and *CREBBP* (11%), which are associated with proliferation and cell survival (1) (45) for the former and to chromatin-modification processes (1) (45) for the latter.

We found mutations in the NOTCH gene family, specifically in the NOTCH1 gene, in 10% of our cases, while mutations in the NOTCH2 gene were observed at a slightly lower frequency of 6%. Under normal and non-altered conditions, the NOTCH family genes have been reported to have an inhibitory or blocking effect on the expression of neuroendocrine genes in SCLC cells (1) (45). We found alterations of NOTCH genes to be distributed across all NAPY subtypes.

In our SCLC cohort, we observed mutation frequencies of 9% for *FAT1*, a tumor suppressor gene, and 8% for *ATRX*, a gene encoding a chromatin remodeling protein. These mutation rates align with findings reported in previous research studies on SCLC (1) (47).

We observed NTRK3, a member of the NTRK family, at a mutation rate of 6%. This gene activates different signaling pathways, including the PI3K/AKT and the MAPK pathways, that control cell survival and differentiation. Alterations in *NTRK3* have already been described for SCLC in other research studies (47).

Our results reveal alterations in the RET (7%) and ROS1 (7%) proto-oncogene as had been previously reported for SCLC (36).

We observed *MYCL* amplifications in 6% of our SCLC samples, primarily associated with SCLC-A subtype, followed by SCLC-N. *MYCL* amplifications are linked to cell cycle progression and cell growth (38) (45). Gene amplifications of *MYC* were observed in 3% of the cases, associated with SCLC-N, SCLC-Y and SCLC-P, and absent in the SCLC-A subtype. Mutually exclusive amplification of MYC family genes (*MYC*, *MYCL*, and *MYCN*) has also been reported in other SCLC studies, showing similar subtype associations to our findings (1) (6). Our results show consistency with higher expression of the MYC hallmark50 signature in SCLCs with *MYC* or *MYCL* amplification compared to wild type SCLCs (Figure 25).

The investigation of DNA-level alterations across our 332 SCLCs confirms that the NAPY expression subtypes are an alternative classification scheme for SCLC complementing subtyping by genomic alterations.

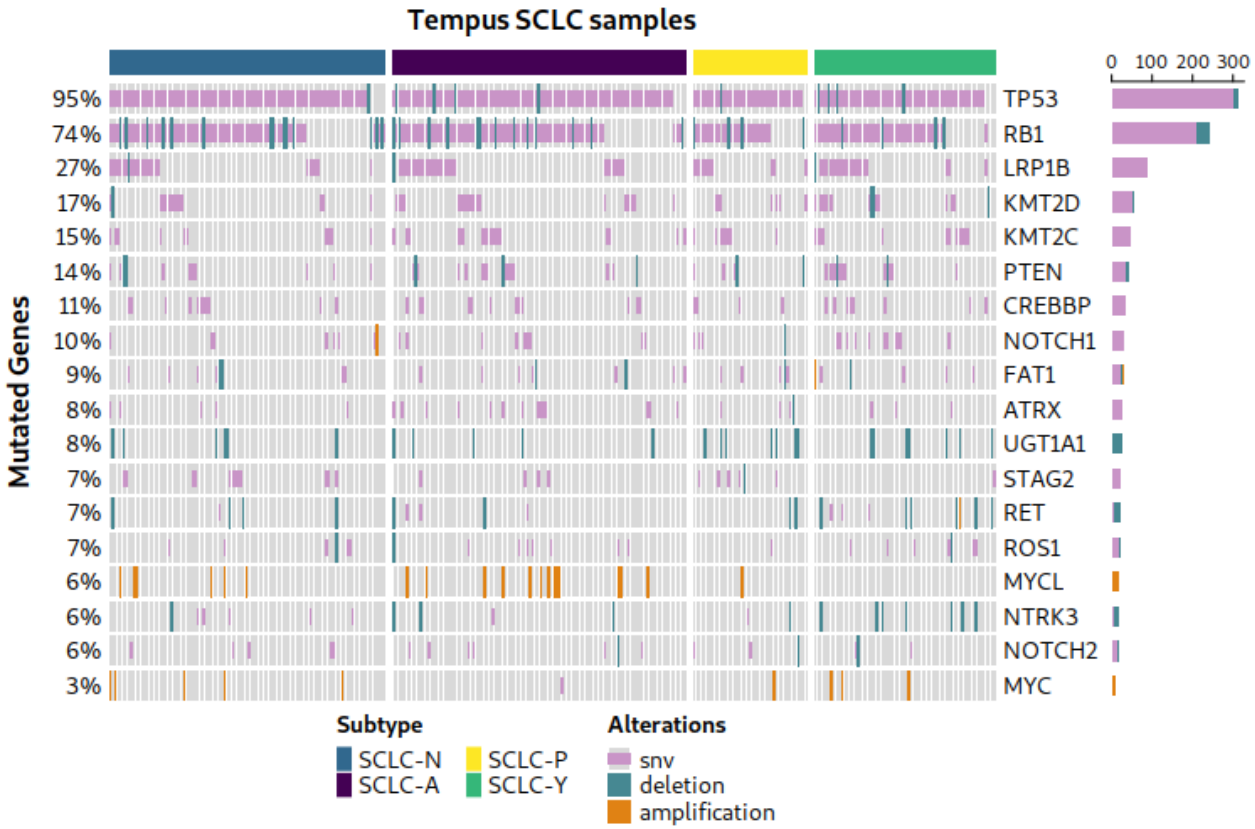

Figure 24. Oncoprint of genomic alterations across subtypes for the 332 SCLC tumor samples from Tempus with a frequency greater than 5%. Alterations in *MYC* gene are included for informational purposes. Single nucleotide variants in pink, copy number deletions in blue and copy number amplifications in orange. Copy number variants were reported in samples with observed copy number gains of eight or more copies, or when a homozygous loss was detected.

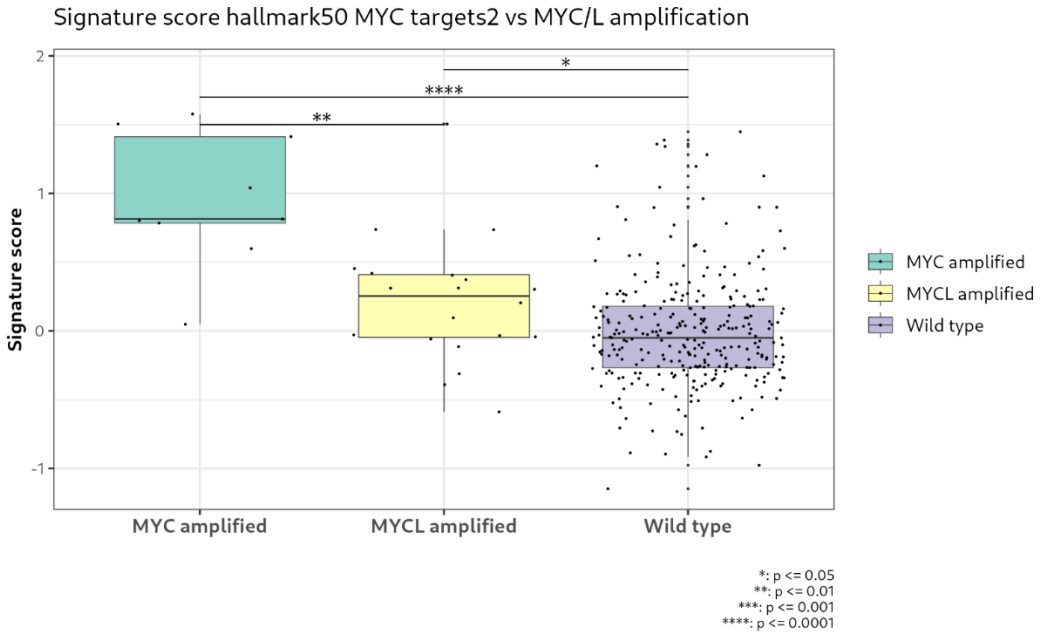

Figure 25. Boxplot of MYC signature score (hallmark50 targets 2 Liberzon *et al.* vs samples with MYC amplified, MYCL amplified and wild type MYC/L.

6.6 Outcome assessment results

Analysis of clinical data of the Tempus cohort (Table 10) reveals that most of the patients received chemotherapy in combination with immunotherapy checkpoint inhibitors as the first lot of treatment.

1st Lot	Overall, N = 307[1]	A, N = 102[1]	N, N = 99[1]	P, N = 42[1]	Y, N = 64[1]
Treatment received					
Biologic + Chemotherapy + Immunotherapy (IO) Checkpoint Inhibitor	1 (0.3%)	1 (1.0%)	0 (0%)	0 (0%)	0 (0%)
Chemotherapy	127 (41%)	46 (45%)	39 (39%)	15 (36%)	27 (42%)
Chemotherapy + Immunotherapy (IO) Checkpoint Inhibitor	175 (57%)	55 (54%)	59 (60%)	26 (62%)	35 (55%)
Immunotherapy (IO) Checkpoint Inhibitor	3 (1.0%)	0 (0%)	0 (0%)	1 (2.4%)	2 (3.1%)
Tyrosine Kinase Inhibitors (TKI)	1 (0.3%)	0 (0%)	1 (1.0%)	0 (0%)	0 (0%)

[1] n (%)

Table 10. Summary of first lot of treatment received on Tempus SCLC patients.

The preference for chemotherapy in combination with immunotherapy over chemotherapy treatment alone may be attributed to timing of initial treatment, with most patients in the cohort receiving their first treatment during or after 2019. Figure 26 illustrates the temporal evolution of treatment selection for the Tempus SCLC cohort. Notably, the combination of chemotherapy with immunochemotherapy became the new standard of care for first-line treatment of extensive-stage SCLC in 2019 (48) (49). Figure 27 shows the temporal evolution of treatment selection for the Tempus SCLC cohort with Stage 4.

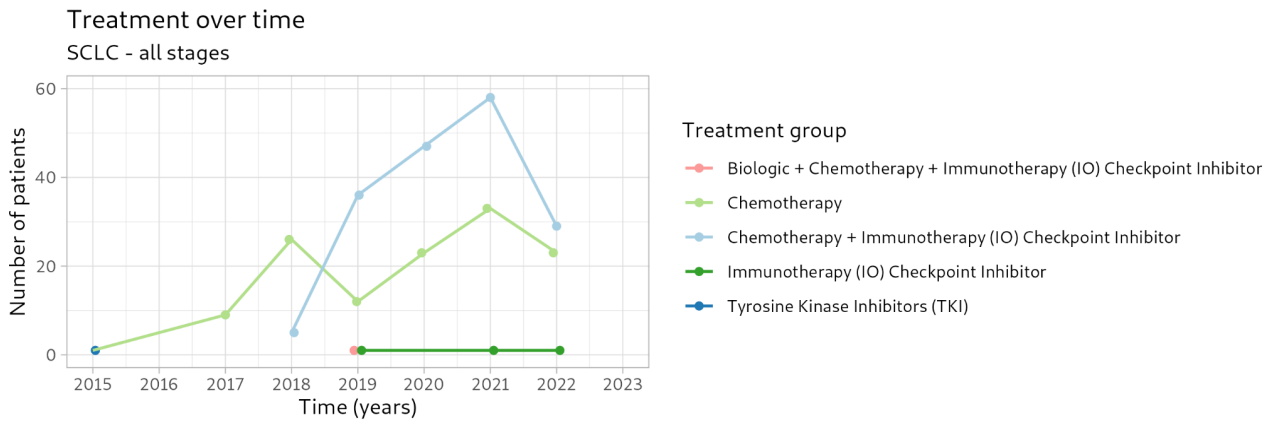

Figure 26. First lot of treatment received over time for the Tempus SCLC cohort (all stages of disease)

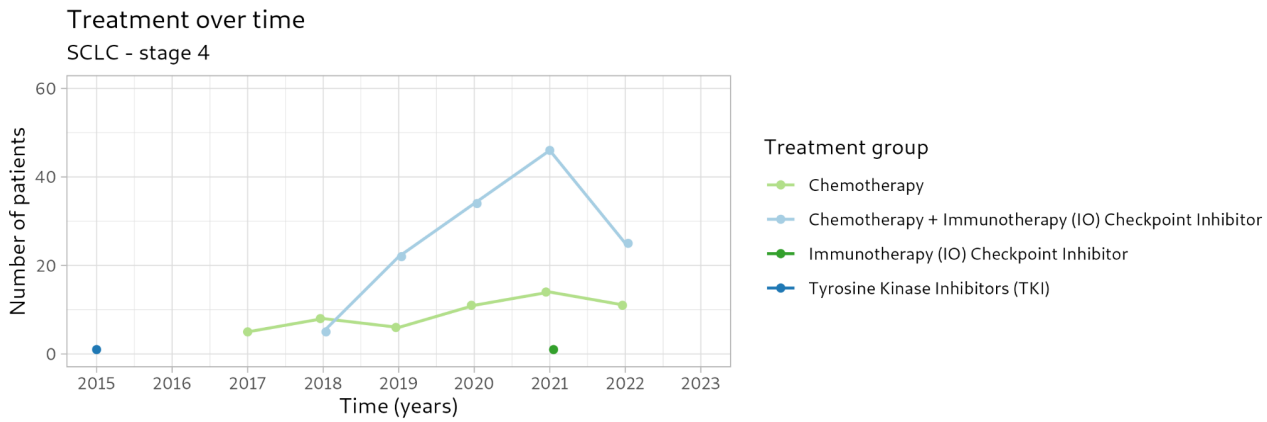

Figure 27. First lot of treatment received over time for the Tempus SCLC cohort (stage 4)

The overall survival analysis for the 207 SCLC Tempus patients with stage 4 of disease at time of diagnosis, over a 1-year evaluation period is shown in Figure 28, Table 11.
All regression coefficients have p-values higher than 0.05, so the null hypothesis stating the molecular subtype does not have an influence on the probability of survival cannot be rejected.

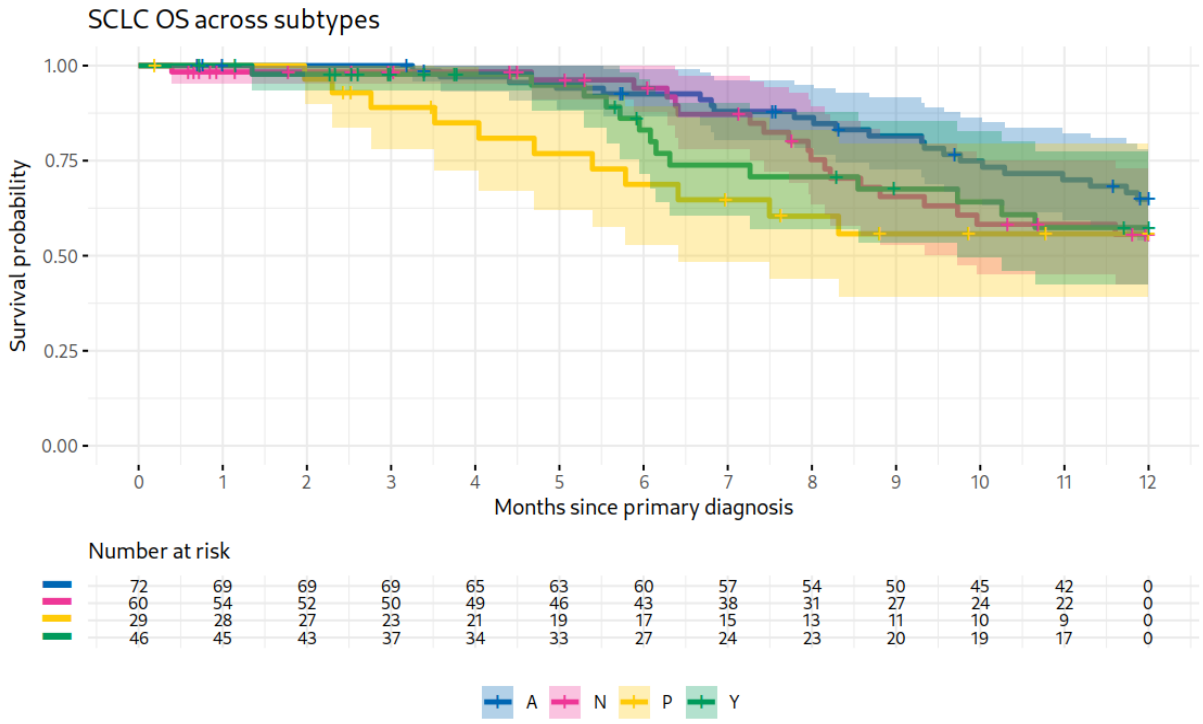

Figure 28. Overall Survival after 12 months for the 207 SCLC patients in stage 4 at time of diagnosis, compared by molecular subtype.

Characteristic	HR[1]	95% CI[1]	p-value
subtype			
A	—	—	
N	1.37	0.74, 2.54	0.3
P	1.86	0.90, 3.85	0.092
Y	1.41	0.72, 2.75	0.3

[1] HR = Hazard Ratio, CI = Confidence Interval

Table 11. Coefficients of the hazard regression function with their statistical significance, taking subtype A as reference for calculating the hazard ratio, related to Figure 22 OS analysis curve.

However, since the treatment received and the response to treatment can be variables influencing the probability of survival, and therefore biasing the results, the overall survival analysis was restricted to the 132 patients that received chemotherapy in combination with immunotherapy as first line of treatment and studied over 9 months (Figure 29, Table 12).

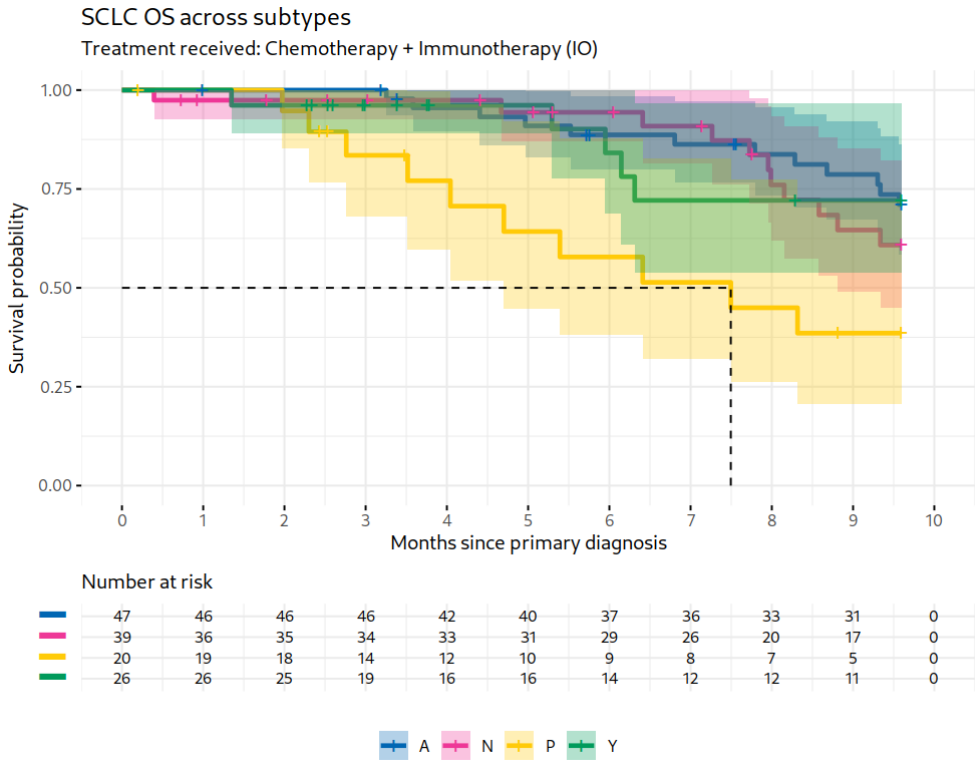

Figure 29. Overall Survival after 9 months for the 132 SCLC patients in stage 4 at time of diagnosis, who received chemotherapy in combination with immunotherapy as the first batch of treatment, compared by molecular subtype.

Characteristic	HR[1]	95% CI[1]	p-value
subtype			
A	—	—	
N	1.34	0.59, 3.03	0.5
P	3.44	1.48, 7.99	0.004
Y	1.09	0.38, 3.09	0.9

[1] HR = Hazard Ratio, CI = Confidence Interval

Table 12. Coefficients of the hazard regression function with their statistical significance, taking subtype A as reference for calculating the hazard ratio, related to Figure 23 OS analysis curve.

While SCLC-Y has been suggested to be particularly susceptible to immunotherapy and more likely to respond to immune checkpoint inhibitors due to the upregulation of PD-L1 transcript by YAP1 and higher expression levels of CD38 in SCLC-Y tumor cells (17) (19) our overall survival analysis did not reveal a significantly better survival probability for the SCLC-Y compared to the other subtypes. Our findings indicate that SCLC-P patients have a significantly lower probability of survival compared to SCLC-A patients (HR: 3.44, p-value: 0.004) over a 9-month period when diagnosed at stage 4 and treated with chemotherapy in combination with immunotherapy. Furthermore, SCLC-P was the only group to experience a decrease in the probability of survival to 50%, highlighting the poorest survival outcomes among all subtypes.

Nevertheless, a more comprehensive analysis is needed to draw clinical conclusions, taking into account other variables that may also influence survival probability and introduce bias in the results, such as gender, sex, smoking status, node involvement, etc.

6.7 Discussion Summary

Despite the presence of characteristic genomic alterations in SCLC, no exclusive mutations were observed across subtypes. Common alterations in *TP53, RB1, LRP1B, KMT2D,* and *KMT2C* were found in all subtypes. Less frequently altered genes such as MYCL, while appearing less affected in SCLC-A, were not entirely absent in this subtype. In this study, we validated in 332 new SCLC samples the heterogeneous mutational landscape previously described in other papers (45) and therefore, emphasize the significance of gene expression in defining distinct molecular subtypes.

Gene expression signatures are a powerful tool to understand and characterize the different molecular characteristics of the SCLC subtypes. In this study, we have evaluated and further validated in 332 SCLC samples the differential expression of relevant biological signatures across subtypes.

For instance, our results demonstrate a remarkably more immune phenotype for the SCLC-Y subtype, with differential expression of interferon and macrophage signatures compared to the other subtypes. This characteristic inflammatory profile with immune cell infiltration has been extensively discussed in other studies (15) (16).

Moreover, our results on the epithelial to mesenchymal transition signature provide additional evidence on the remarkably mesenchymal phenotype of the SCL-Y subtype. Although SCLC is considered an epithelial malignancy, other studies have also demonstrated that SCLC-Y is the most mesenchymal subtype and have suggested this feature as a possible mechanism of resistance (15).

In the context of SCLC disease, *NOTCH* and *MYC* have also been described as relevant biological pathways. Several research groups have demonstrated that NOTCH pathway inactivation or downregulation is essential for the neuroendocrine phenotypes (SCLC-A and SCLC-N) (9) (28) (38) and that, conversely, SCLC-Y and SCLC-P subtypes show a positive correlation with NOTCH genes (38). Similar behavior has been described for the MYC signature by these research groups, consisting of a negative correlation between MYC and the neuroendocrine score, and a positive and preferentially high correlation with the SCLC-P subtype (9) (28).

Our results are in alignment with previous findings, exhibiting differential expression of the NOTCH signature in the SCLC-Y and SCLC-P subtypes, and showing the highest expression of the MYC signature in the SCLC-P subtype. It is important to note, however, that the NOTCH expression signature evaluated has been specifically developed in the context of breast cancer.

Although many of the expression signatures evaluated showed significant differences between subtypes, not all subtypes were successfully differentiated by a unique signature. For instance, the neuroendocrine signature exhibited differential expression between the neuroendocrine phenotypes (SCLC-A/N) and the non-neuroendocrine phenotypes (SCLC-P/Y), but couldn't differentiate between SCLC-A and SCLC-N. Conversely, the macrophage signature exhibited differential expression between the non-neuroendocrine phenotypes (SCLC-P/Y) and the neuroendocrine phenotypes (SCLC-A/N), but couldn't distinguish between SCLC-P and SCLC-Y.

In this study, we developed four subtype-specific downstream programs with sufficient signaling power to differentiate each of the four SCLC subtypes and repredict the classes. Each of our subtype-specific signatures comprises 20 genes, many of which have been previously identified as differentially expressed in a specific subtype by other research groups (5) (18). Gene ontology analysis of these downstream programs revealed enrichment in TGFB, interferon-alpha, and Hippo pathway terminologies within our YAP1-associated expression signature, aligning with the results from our differential expression analysis across subtypes. The results for the WNT pathway in the SCLC-Y subtype were inconclusive; our YAP1-associated signature showed enrichment in terms related to negative regulation, yet the WNT expression signatures evaluated in the SCLC-Y samples displayed high scores compared to other subtypes. Both the Hippo and WNT signaling pathways

are pivotal in maintaining tissue homeostasis and regulating organ size, primarily through their roles in cell proliferation, differentiation, and apoptosis, and are frequently seen dysregulated in human cancers (50). While the crosstalk between these two pathways has been extensively studied (51) (52), the specific contributions and potential overlapping roles in SCLC-Y samples warrant further investigation. Genes within our *ASCL1*- and *NEUROD1*- expression signatures exhibited enrichment in ontology terms related to neuropeptide signaling and synapses. Specifically, dopaminergic synapses were associated with the *ASCL1* expression signature, while cholinergic synapses were linked to the *NEUROD1* expression signature. For our *POU2F3*- gene expression signature, gene ontology revealed an association with negative regulation of Schwann cell proliferation, negative regulation of *EGFR* pathway and of low-density lipoprotein clearance. While *Cao et al. (2022)* demonstrated the crosstalk between SCLC and tumor-associated Schwann cells, exploring potential differences between subtypes, particularly with SCLC-P, would be of interest.

Our findings also demonstrate the potential of machine learning classifiers based on transcriptomic data, particularly on subtype-specific downstream programs, as a viable approach for addressing the SCLC subtype classification problem. Notably, our study did not identify a single algorithm that significantly outperformed the others. Furthermore, while the optimal number of features may vary slightly depending on the model, this variation did not translate into more than 5% of performance improvement. In our study, we have selected a support vector machine classifier trained on 4 subtype-specific downstream programs of 20 genes each to run the complete pipeline and achieved an accuracy of 90%. While these results demonstrate the power of our subtype-specific downstream programs, we do not rule out the possibility of achieving further improvement with the inclusion of additional genes or complementary expression signatures. Moreover, evaluating the classifier on new datasets is advisable to further extend the validation.

8. Conclusion

In this study, we demonstrate the presence of reliable multi-gene expression signatures linked to the four key transcription factors *NEUROD1, ASCL1, POU2F3 and YAP1,* which exhibit sufficient signaling to distinguish the four molecular subtypes of SCLC and replicate the classes with a 90% accuracy when fed to a machine learning classifier. Consequently, these signatures can provide supplementary support for SCLC subtype assignment, offering an alternative to relying solely on the gene or protein expression of the four key transcription factors.

Furthermore, our transcription-factor downstream programs exhibited robustness and can provide an additional layer of characterization, complementing other established signatures evaluated in this study, as evidenced by distinct expression patterns observed across subtypes.

These results also aim to encourage the utilization and further development of computational models, such as machine learning techniques, to bolster the classification of SCLC subtypes, with the ultimate goal of establishing a standardized process.

While our analysis of overall survival did not reveal an improved prognosis for SCLC-Y compared to other SCLC subtypes when treated with a combination of chemotherapy and immunotherapy, it's important to note certain limitations in the study cohort. These limitations include relatively small sample sizes, consideration of only the first line of treatment, and evaluation of immune checkpoint inhibitors limited to PD-L1 or PD-1 inhibitors. However, our efforts to establish a consensus on the diagnostic procedures for determining SCLC molecular subtypes will pave the way for the design and development of clinical studies aimed at comparing survival outcomes across subtypes with distinct targeted therapeutic approaches

9. Supplementary Information

The clinical and molecular profiling data analyzed in this study were part of the real-world multi-omics cancer database assembled by Tempus AI, Inc. This data is subject to controlled access for privacy and proprietary reasons.

Name	Type of sample	Source	Subtype assignment
COLO-668	Cell line	CCLE	SCLC-A
COR-L47	Cell line	CCLE	SCLC-A
COR-L88	Cell line	CCLE	SCLC-A
COR-L95	Cell line	CCLE	SCLC-A
DMS-153	Cell line	CCLE	SCLC-A
DMS-454	Cell line	CCLE	SCLC-A
DMS-53	Cell line	CCLE	SCLC-A
DMS-79	Cell line	CCLE	SCLC-A
NCI-H1092	Cell line	CCLE	SCLC-A
NCI-H1105	Cell line	CCLE	SCLC-A
NCI-H1184	Cell line	CCLE	SCLC-A
NCI-H1436	Cell line	CCLE	SCLC-A
NCI-H146	Cell line	CCLE	SCLC-A
NCI-H1618	Cell line	CCLE	SCLC-A
NCI-H1836	Cell line	CCLE	SCLC-A
NCI-H1876	Cell line	CCLE	SCLC-A
NCI-H1930	Cell line	CCLE	SCLC-A
NCI-H1963	Cell line	CCLE	SCLC-A
NCI-H2029	Cell line	CCLE	SCLC-A
NCI-H2081	Cell line	CCLE	SCLC-A
NCI-H209	Cell line	CCLE	SCLC-A
NCI-H2196	Cell line	CCLE	SCLC-A
NCI-H510	Cell line	CCLE	SCLC-A
NCI-H69	Cell line	CCLE	SCLC-A
NCI-H889	Cell line	CCLE	SCLC-A
SHP-77	Cell line	CCLE	SCLC-A
COR-L24	Cell line	CCLE	SCLC-N
COR-L279	Cell line	CCLE	SCLC-N
DMS-273	Cell line	CCLE	SCLC-N
HCC-33	Cell line	CCLE	SCLC-N
NCI-H1694	Cell line	CCLE	SCLC-N
NCI-H2171	Cell line	CCLE	SCLC-N
NCI-H2227	Cell line	CCLE	SCLC-N
NCI-H446	Cell line	CCLE	SCLC-N
NCI-H524	Cell line	CCLE	SCLC-N
NCI-H82	Cell line	CCLE	SCLC-N
SCLC-21H	Cell line	CCLE	SCLC-N
COR-L311	Cell line	CCLE	SCLC-P
NCI-H1048	Cell line	CCLE	SCLC-P
NCI-H211	Cell line	CCLE	SCLC-P
NCI-H526	Cell line	CCLE	SCLC-P
DMS-114	Cell line	CCLE	SCLC-Y
NCI-H1341	Cell line	CCLE	SCLC-Y
NCI-H196	Cell line	CCLE	SCLC-Y
NCI-H2286	Cell line	CCLE	SCLC-Y
NCI-H841	Cell line	CCLE	SCLC-Y
SBC-5	Cell line	CCLE	SCLC-Y
SW-1271	Cell line	CCLE	SCLC-Y

Supplementary Table 1. 48 CCLE SCLC Cell lines identification with their molecular label.

Metastatic stages	
M0	Indicates that no evidence of metastasis has been detected at the time of diagnosis.
M1	Indicates the presence of metastasis, which means the spread of cancer to other distant parts of the body.
MX	Is used when there is insufficient information available to determine the presence or absence of metastasis
Node Involvement	
N0	Indicates no regional lymph node involvement. No cancer cells are found in the nearby lymph nodes.
N1-N3	These categories represent increasing degrees of regional lymph node involvement, where N1 indicates involvement of nearby lymph nodes, and N2 and N3 indicate involvement of lymph nodes further away or in multiple regions.
NX	Denotes that lymph node involvement cannot be assessed or the information is not available.
Stages	
0	This stage is typically referred to as carcinoma in situ. It indicates abnormal cells that are present only in the layer of cells where they first developed and have not spread to nearby tissues.
1	This stage generally denotes early cancer that is contained within the organ or tissue of origin. It may indicate a small tumor size with minimal or no spread to nearby lymph nodes or other organs.
2	This stage usually implies a larger tumor size or increased involvement of nearby tissues. It may also suggest the presence of cancer cells in nearby lymph nodes.
3	This stage often indicates more extensive local spread of cancer to nearby structures or tissues. It may also involve significant lymph node involvement, indicating regional spread.
4	This stage typically represents advanced cancer that has spread to distant organs or tissues. It often indicates the presence of metastasis, which can occur via the bloodstream, lymphatic system, or direct invasion.
Tumor Grades	
T1	Indicates a small tumor that is confined to the tissue or organ of origin. It typically implies a tumor size below a specific threshold.
T2	Represents a larger tumor compared to T1, usually with an increased size or extent of invasion into surrounding tissues.
T3	Denotes a locally advanced tumor that has further invaded nearby structures, organs, or tissues beyond its site of origin. The exact characteristics of T3 may vary depending on the specific cancer type.
T4	Indicates an even more extensive tumor that has invaded adjacent structures, organs, or tissues even more extensively compared to T3.

Supplementary Table 2. Oncology definitions, abbreviations and nomenclature related to Table 4.

	G1				G2				G3				G4				G5			
	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1
1	DDC	CERKL	C11orf53	LATS2	DDC	CERKL	C11orf53	WWTR1	DDC	CERKL	C11orf53	WWTR1	DDC	CERKL	C11orf53	LATS2	DDC	CERKL	C11orf53	WWTR1
2	SEC11C	SSTR2	GFI1B	CYBRD1	CNKSR3	SSTR2	TRPM5	LATS2	SEC11C	SSTR2	APOBEC1	LATS2	SEC11C	SSTR2	GFI1B	WWTR1	SEC11C	SSTR2	TRPM5	LATS2
3	CNKSR3	CHRNB4	TRPM5	WWTR1	SEC11C	NHLH1	APOBEC1	MSRB3	RIMKLA	NHLH1	FAM150A	OSMR	SCN3A	NHLH1	APOBEC1	OSMR	PTPRN2	NHLH1	GFI1B	OSMR
4	PTPRN2	CHRNA3	APOBEC1	OSMR	RIMKLA	CHRNB4	GFI1B	LAMB2	GRP	CHRNA3	TRPM5	LAMB2	CNKSR3	CHRNB4	TRPM5	LAMB2	RIMKLA	CHRNB4	APOBEC1	LAMB2
5	SCN3A	NHLH1	FAM150A	MSRB3	RGS17	NEUROD4	PTPN18	OSMR	STK32A	CHRNB4	GFI1B	CYBRD1	PTPRN2	NEUROD4	FAM150A	EHD2	CNKSR3	PPP1R17	FAM150A	CYBRD1
6	RIMKLA	NEUROD4	PTPN18	MRC2	MS4A8	CHRNA3	IMP4	CYBRD1	CNKSR3	NEUROD4	BMX	RBMS3	RIMKLA	PPP1R17	PTPN18	CYBRD1	RGS17	CHRNA3	ART3	GPX8
7	SMPD3	PPP1R17	LANCL3	LAMB2	SCN3A	LMO1	FAM150A	LIMS2	PTPRN2	LMO1	PTPN18	GPX8	RGS17	CHRNA3	HES2	MSRB3	GRP	THSD7B	PTPN18	MSRB3
8	STK32A	LMO1	ART3	GPX8	PTPRN2	PPP1R17	ART3	EHD2	RGS17	PPP1R17	LANCL3	MSRB3	STK32A	LMO1	IMP4	MYL9	SCN3A	LMO1	IMP4	EPHA2
9	RGS17	SHF	BMX	SYDE1	STK32A	THSD7B	FOXI1	SYDE1	SMPD3	SHF	IMP4	WTIP	GRP	THSD7B	BMX	SYDE1	STK32A	NEUROD4	BMX	SYDE1
10	NOL4	SLC17A6	HSPA8P4	MYL9	FOXA2	DACH1	CALHM3	PTRF	FBLN7	FNDC5	ADAMTS19	PMP22	ETS2	SLC17A6	CALML5	GPX8	CACNA1A	SHF	CALHM3	PMP22
11	CA8	THSD7B	HES2	PMP22	SMPD3	HPCA	ASCL2	GPX8	CACNA1A	NHLH2	ART3	PTRF	MS4A8	CNTN2	LRMP	ITGB5	MS4A8	SLC17A6	VSNL1	ITGB5
12	MS4A8	PLCL2	IMP4	EHD2	CA8	NEUROD2	VSNL1	PMP22	SCN3A	KIAA1614	LRMP	SYDE1	ICA1	DACH1	ASCL2	LIMS2	NOL4	FNDC5	HES2	EHD2
13	GRP	NEUROD2	RGS13	AXL	GRP	FNDC5	EN1	HEPH	MS4A8	CNTN2	RGS13	FGD5	NOL4	NEUROD2	C1orf61	PTRF	SMPD3	NHLH2	ASCL2	PLA2R1
14	CAMK1D	RGS8	VSNL1	WTIP	SCNN1A	SHF	KLHDC7A	MYL9	ETS2	THSD7B	HSPA8P4	LIMS2	CA8	IGDCC3	ART3	IFITM3	CNTNAP2	DACH1	LRMP	MYL9
15	SCNN1A	PDZRN4	SH2D6	PTRF	ICA1	SLC17A6	PVRL4	WTIP	NR0B2	DACH1	CALHM3	IL1R1	CACNA1A	SHF	IL19	WTIP	CA8	NEUROD2	HSPA8P4	WTIP
16	CACNA1A	SCN1B	KLHDC7A	LIMS2	RAB3B	SCN1B	ANO7	ITGB5	CA8	CAMKV	ANO7	EHD2	SCNN1A	FNDC5	LANCL3	PMP22	ETS2	KIAA1614	IL19	ZCCHC24
17	NR0B2	GNG8	LRMP	ITGB5	CACNA1A	CNTN2	LANCL3	AXL	CAMK1D	HPCA	HES2	AXL	NR0B2	CLVS1	SH2D6	IL1R1	RAB3B	GNG8	RGS13	MRC2
18	ETS2	KIAA1614	C1orf61	IL1R1	NOL4	PTCHD2	HES2	ZCCHC24	ICA1	NEUROD2	IL19	NOTCH3	SMPD3	HPCA	KLHDC7A	HEPH	SLC36A4	CNTN2	ADAMTS19	AXL
19	DLL3	EBF3	CALHM3	VIM	ETS2	IGDCC3	COLCA2	FGD5	SLC36A4	SEMA6A	VSNL1	HEPH	SCN2A	SCN1B	RGS13	MRC2	NR0B2	IGFBPL1	LANCL3	PTRF
20	CNTNAP2	TCP10L	COLCA2	NOTCH3	NR0B2	GKAP1	C1orf61	VIM	RAB3B	CLVS1	SH2D6	CAV2	CNTNAP2	KIAA1614	CALHM3	AXL	CERS4	GAS2	KLHDC7A	LIMS2
21	RGS7	CNTN2	ANO7	IFITM3	CNTNAP2	NHLH2	HSPA8P4	RBPMS	KIAA1211L	SLC17A6	MOCOS	PLA2R1	RAB3B	RBFOX3	ANO7	EPHA2	SCNN1A	IGDCC3	FAM124A	IL1R1
22	ICA1	FNDC5	ASCL2	RHOC	CAMK1D	PDE1C	OBP2B	EPHA2	SCNN1A	EBF3	UGT2B28	MYL9	FOXA2	PDE1C	ALDH3B2	PLA2R1	ICA1	SCN1B	C1orf61	CAV2
23	FBLN7	GAS2	PVRL4	TLCD2	DLL3	NTN3	MOCOS	CRTAP	NOL4	IGDCC3	ASCL2	ZCCHC24	FBLN7	GAS2	EN1	COL5A2	RIMS2	HPCA	MOCOS	RBMS3
24	NELL1	DACH1	FOXI1	EFEMP2	NELL1	KIAA1614	CALML5	ENG	NKX2-1	GKAP1	KIAA1024L	CRIM1	CAMK1D	PLCL2	FOXI1	ZCCHC24	PRUNE2	CAMKV	ALDH3B2	IFITM3
25	NKX2-1	GKAP1	MYB	HEPH	SLC36A4	PRIMA1	MYB	MRC2	PRUNE2	SCN1B	FAM124A	CAV1	ERO1LB	NHLH2	FAM124A	RBMS3	FBLN7	PTCHD2	PVRL4	FGD5
26	SLC36A4	NHLH2	ALDH3B2	SERPINH1	PRUNE2	GAS2	FAM124A	RBMS3	SERGEF	PRDM8	KLHDC7A	RBPMS	SLC36A4	MMD2	PVRL4	CCND1	ERO1LB	UMODL1	FOXI1	RHOC
27	CXXC4	PROKR1	FAM124A	FGD5	USP41	CLVS1	BMX	DCN	ENTPD8	FAM71C	SOSTDC1	IFITM3	WNT11	GNG8	COLCA2	THBS1	CAMK1D	GKAP1	ANO7	TLCD2
28	PRUNE2	HUNK	BARX2	RBMS3	ST18	CAMKV	IL19	VEGFC	CERS4	NTN3	ALDH3B2	CARD6	NKX2-1	PROKR1	UGT2B28	SERPINH1	FOXA2	EBF3	EN1	RAB3IL1
29	RAB3B	TSPAN18	IL19	COL5A2	DGKB	PROKR1	LRMP	CAV2	WNT11	RBFOX3	KCNQ4	MRC2	PRUNE2	PDZRN4	VSNL1	COL12A1	PCLO	PLCL2	BARX2	NOTCH3
30	CADPS2	RBFOX3	ATL2	CAV1	ERO1LB	FRMD3	EFNA4	CAV1	ZMAT4	IGFBPL1	PVRL4	SHROOM4	RGS7	TSPAN18	MYB	PPIC	WNT11	RBFOX3	UGT2B28	THBS1
31	LFNG	HPCA	TULP1	EPHA2	WNT11	PDZRN4	KCTD1	IL1R1	TMEM150C	MDGA1	C1orf61	MRGPRF	DLL3	CAMKV	EFNA4	RHOC	DLL3	RGS8	MYB	SERPINH1
32	WNT11	PRDM8	CALML5	CTGF	CADPS2	GNG8	BARX2	COL5A2	TFF3	PLCL2	ADSS	EDNRA	ZMAT4	TCHH	FCHSD2	NOTCH3	UNC13A	ATP2B2	CALML5	CAV1
33	ERO1LB	PTCHD2	RNF223	PLA2R1	SCGN	TCP10L	ADAMTS19	IFITM3	NELL1	TCP10L	COLCA2	ITGB5	ELOVL7	PRDM8	OBP2B	CAV2	USP41	PRIMA1	PRSS21	CRIM1
34	DNALI1	IGDCC3	EN1	RAB3IL1	ZMAT4	HUNK	ALDH3B2	COL6A2	USP41	ATP2B2	PRSS21	FBLN5	CADPS2	SEMA6A	RNF223	CAV1	NKX2-1	HUNK	SH2D6	HEPH
35	CERS4	FRMD3	GALNT14	DDR2	FAM155A	RGS8	RASD2	COL6A3	CXXC4	PDZRN4	FOXI1	ARHGAP31	ST18	NTNG2	MOCOS	MYOF	UNC80	PDZRN4	PNOC	MYOF
36	ZMAT4	PGF	ADSS	CAV2	RGL3	SEMA6A	FAM117A	FBLN5	DGKB	NPS	FAM117A	SERPINH1	SERGEF	ATP2B2	B4GALT5	PLD1	ST18	PRDM8	EFNA4	RBPMS
37	RAB19	CAMK4	TAF4B	RBPMS	PCLO	ATP2B2	C12orf74	GPR17	RGS7	GAS2	SLC12A8	PPIC	DGKB	PTCHD2	TAS2R60	FGD5	SERGEF	PDE1C	TULP1	KIF1C
38	SYT4	SEMA6A	MPV17L	CYR61	PCSK1	TTC7B	UGT2B28	EDNRA	DLL3	TTC7B	ADAT2	HMCN1	RNF148	IGFBPL1	HSPA8P4	CTGF	RGS7	CAMK4	SOSTDC1	DDR2
39	SERGEF	NPS	KAZN	ZCCHC24	TMEM178B	IGFBPL1	SLC12A8	DLC1	SEPW1	MMP24	GNAT3	MYOF	CERS4	MDGA1	ADAMTS19	VIM	RGL3	TCP10L	GALNT14	VIM
40	RNF148	BMPER	MOCOS	HEG1	EPOR	MDGA1	PNOC	CD34	PCSK2	PRIMA1	RNF223	EPHA2	EPOR	RGS8	TMPRSS13	GPR17	DGKB	SEMA6A	COLCA2	CRTAP
41	JAM3	PRIMA1	ADAT2	TNFRSF10B	RAB19	FAM71C	B4GALT5	RAB3IL1	SCN2A	GNG8	EFNA4	ENG	USP41	PRIMA1	BARX2	HTRA1	EPOR	PROKR1	ADSS	C1QTNF5
42	TFF3	MRAP2	INPP5B	COL6A2	NKX2-1	RCOR2	GPR110	SERPINH1	TMEM61	PTCHD2	IL10	DDR2	SPTB	RCOR2	MPV17L	CYR61	CHGA	FAM212B	KAZN	NNMT
43	FOXA2	FYN	NCOA3	RRAS	CERS4	PLCL2	PIRT	NOTCH3	FAM155A	PDE1C	MPV17L	VIM	RAB19	SYT6	CCDC115	COL6A2	SPTB	GJD2	RNF223	CYR61
44	RNF133	TSHR	SGCG	COL12A1	SPTB	RBFOX3	ADTRP	RHOC	INPPL1	MRAP2	TAF4B	DAB2	KIAA1211L	RNF112	TFAP2C	CARD6	FAM155A	NTN3	TAF4B	CARD6
45	ST18	UMODL1	B4GALT5	CRTAP	FBLN7	RNF112	NREP	PDGFRB	CNTNAP2	BARHL1	MYB	PLD1	FAM155A	FRMD3	RBM38	CD93	CXXC4	CORO2B	ATL2	CCDC102A
46	SCN2A	MDGA1	EFNA4	ENG	RNF148	PRDM8	ACSS1	EMILIN1	SPTB	UMODL1	ATL2	LAMA4	GCH1	GKAP1	FAM117A	C1QTNF5	RUNDC3A	CLVS1	OBP2B	REST
47	RET	PDE1C	ACSS1	CRIM1	SERGEF	FYN	ACADSB	DDR2	RGL3	FRMD3	ANXA1	PLAT	RGL3	GRIN2D	KAZN	EMILIN1	TFF3	NPS	MPV17L	GPR17
48	RGL3	DCC	KCTD1	DLC1	PCSK2	MMD2	GALNT14	EFEMP2	FOXA2	ASIC1	FCHSD2	SLC2A10	SMOC2	KCNQ2	SLC12A8	DAB2	SCN2A	BMPER	KCNQ4	TFPI
49	SMOC2	EPHB2	C12orf74	GPR17	HABP2	ZNF462	MPV17L	PLAT	DBH	CADPS	RGS21	C1QTNF5	PCSK1	GJD2	ATL2	CD34	PCSK1	FRMD3	IL10	EDNRA
50	KIAA1211L	LZTS1	KCNQ4	EDNRA	RNF133	ANKS1B	PLEKHG7	CRIM1	EPOR	DENND2A	CTCFL	KDR	CXXC4	TTC7B	NCOA3	RBPMS	SCG2	CADPS	FCHSD2	SLC2A10

Supplementary Table 3. 80 predictor groups evaluated in the machine learning pipeline (from group 1 to group 5). The purple dashed line separates the 10-gene, 20-gene- 30-gene, 50-gene length signatures.

	G6				G7				G8				G9				G10			
	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1
1	DDC	CERKL	C11orf53	LATS2	SEC11C	CERKL	C11orf53	LATS2	DDC	CERKL	C11orf53	WWTR1	SEC11C	CERKL	C11orf53	LATS2	DDC	CERKL	C11orf53	WWTR1
2	SEC11C	SSTR2	FAM150A	WWTR1	DDC	SSTR2	GFI1B	WWTR1	SEC11C	SSTR2	GFI1B	LATS2	DDC	SSTR2	APOBEC1	OSMR	SEC11C	SSTR2	APOBEC1	LATS2
3	CNKSR3	CHRNB4	APOBEC1	OSMR	CNKSR3	NHLH1	APOBEC1	OSMR	RGS17	NHLH1	TRPM5	OSMR	CNKSR3	CHRNB4	GFI1B	WWTR1	RIMKLA	NEUROD4	GFI1B	OSMR
4	PTPRN2	NEUROD4	TRPM5	LAMB2	RIMKLA	CHRNB4	TRPM5	MSRB3	CNKSR3	NEUROD4	APOBEC1	LAMB2	RIMKLA	NHLH1	FAM150A	LAMB2	RGS17	NHLH1	TRPM5	CYBRD1
5	RIMKLA	CHRNA3	GFI1B	CYBRD1	SCN3A	NEUROD4	PTPN18	EHD2	RIMKLA	CHRNA3	FAM150A	MSRB3	PTPRN2	CHRNA3	TRPM5	CYBRD1	CNKSR3	CHRNB4	PTPN18	MSRB3
6	RGS17	NHLH1	ART3	GPX8	STK32A	CHRNA3	BMX	CYBRD1	SCN3A	CHRNB4	PTPN18	CYBRD1	SCN3A	NEUROD4	BMX	GPX8	PTPRN2	CHRNA3	HSPA8P4	ITGB5
7	STK32A	LMO1	PTPN18	MSRB3	SCNN1A	PPP1R17	FAM150A	LAMB2	GRP	LMO1	LANCL3	EHD2	GRP	LMO1	PTPN18	MSRB3	STK32A	PPP1R17	IMP4	GPX8
8	SCN3A	PPP1R17	IMP4	PMP22	ETS2	LMO1	IMP4	MYL9	PTPRN2	PPP1R17	IMP4	HEPH	CACNA1A	PPP1R17	LRMP	EPHA2	SMPD3	LMO1	FAM150A	LAMB2
9	CA8	SLC17A6	ANO7	SYDE1	MS4A8	THSD7B	LANCL3	SYDE1	STK32A	THSD7B	FOXI1	SYDE1	RGS17	SLC17A6	LANCL3	PMP22	GRP	NEUROD2	ART3	EHD2
10	GRP	THSD7B	LANCL3	WTIP	GRP	CNTN2	KLHDC7A	PTRF	SMPD3	NEUROD2	BMX	MYL9	STK32A	THSD7B	CALHM3	SYDE1	SCN3A	THSD7B	VSNL1	WTIP
11	SMPD3	SHF	HES2	PTRF	PTPRN2	DACH1	HES2	LIMS2	CACNA1A	SHF	CALHM3	LIMS2	CAMK1D	SHF	IQCJ	EHD2	NOL4	SHF	COLCA2	LIMS2
12	CAMK1D	NEUROD2	RGS13	AXL	RGS17	SHF	IQCJ	GPX8	MS4A8	FNDC5	ASCL2	WTIP	NOL4	NEUROD2	IMP4	AXL	ICA1	SLC17A6	ANO7	PMP22
13	MS4A8	DACH1	COLCA2	MYL9	SMPD3	NEUROD2	ART3	PMP22	NR0B2	SLC17A6	LRMP	GPX8	MS4A8	CNTN2	ART3	WTIP	DLL3	CNTN2	HES2	CAV2
14	NR0B2	CNTN2	VSNL1	LIMS2	ICA1	FNDC5	C1orf61	AXL	PRUNE2	CNTN2	HES2	PMP22	ETS2	DACH1	SH2D6	MYL9	NR0B2	NHLH2	CALML5	HEPH
15	NOL4	HPCA	CALHM3	PLA2R1	NR0B2	IGDCC3	LRMP	IFITM3	ETS2	DACH1	ANO7	MRC2	SMPD3	FNDC5	FOXI1	LIMS2	CA8	PLCL2	ADAMTS19	MYL9
16	ICA1	PDZRN4	KLHDC7A	EHD2	CA8	SCN1B	FOXI1	WTIP	RAB3B	HPCA	ART3	ITGB5	CA8	GNG8	C1orf61	PPIC	CAMK1D	FNDC5	CALHM3	PLA2R1
17	FBLN7	KIAA1614	HSPA8P4	HEPH	CACNA1A	CLVS1	ANO7	MRC2	ICA1	NHLH2	COLCA2	AXL	ZMAT4	KIAA1614	HES2	MRC2	SCNN1A	SCN1B	MOCOS	SYDE1
18	CACNA1A	SCN1B	ALDH3B2	IL1R1	CAMK1D	HPCA	SH2D6	HEPH	SCNN1A	KIAA1614	HSPA8P4	PTRF	NKX2-1	CLVS1	FAM124A	IL1R1	FBLN7	IGDCC3	PVRL4	IL1R1
19	ETS2	FNDC5	RNF223	RBMS3	NKX2-1	SLC17A6	CALML5	IL1R1	NOL4	RBFOX3	FAM117A	COL5A2	DGKB	IGDCC3	KLHDC7A	TNFRSF10B	SLC36A4	HPCA	ASCL2	RBMS3
20	ZMAT4	GKAP1	SH2D6	MRC2	CADPS2	KIAA1614	ASCL2	CAV1	CAMK1D	SCN1B	B4GALT5	RBMS3	CERS4	HPCA	CALML5	PTRF	CNTNAP2	SEMA6A	LANCL3	MRC2
21	SLC36A4	IGDCC3	IL19	VIM	NELL1	PLCL2	VSNL1	ITGB5	FBLN7	GKAP1	PVRL4	RBPMS	FBLN7	CAMKV	IL19	IFITM3	RAB3B	DACH1	EN1	CAV1
22	SCNN1A	CLVS1	EN1	IFITM3	RAB19	NHLH2	PVRL4	RBMS3	DLL3	PLCL2	IL19	FGD5	DLL3	SCN1B	ASCL2	RAB3IL1	MS4A8	PROKR1	ALDH3B2	RBPMS
23	KIAA1211L	CAMKV	BMX	FGD5	DLL3	SEMA6A	UGT2B28	VIM	WNT11	TCP10L	RGS13	NOTCH3	ICA1	PDZRN4	RGS13	ITGB5	SERGEF	GKAP1	C12orf74	PTRF
24	RAB3B	RCOR2	PVRL4	SERPINH1	FOXA2	PRDM8	IL19	FGD5	CA8	PDE1C	MPV17L	EPHA2	NR0B2	EBF3	PVRL4	RBMS3	CACNA1A	PRDM8	BMX	AXL
25	FOXA2	FRMD3	LRMP	ZCCHC24	FBLN7	PROKR1	RGS13	COL5A2	SLC36A4	SEMA6A	SH2D6	CDH11	SCNN1A	RBFOX3	MYB	ZCCHC24	NKX2-1	PDE1C	FOXI1	FGD5
26	NKX2-1	SEMA6A	ASCL2	CAV2	NOL4	RGS8	FAM124A	NOTCH3	SERGEF	GAS2	MYB	IL1R1	ERO1LB	NHLH2	ANO7	AJUBA	PCLO	KIAA1614	SOSTDC1	DDR2
27	DLL3	NHLH2	C1orf61	NOTCH3	RNF148	CAMK4	MOCOS	CYR61	CNTNAP2	MDGA1	C1orf61	HTRA1	CNTNAP2	RCOR2	VSNL1	TLCD2	ETS2	PDZRN4	UGT2B28	IFITM3
28	WNT11	RGS8	FAM124A	RBPMS	CNTNAP2	MMD2	HSPA8P4	GPR17	FOXA2	EBF3	CALML5	COL6A2	RGL3	GAS2	ADAMTS19	FGD5	ERO1LB	PTCHD2	FAM117A	NOTCH3
29	NELL1	PLCL2	CALML5	CRIM1	RGS7	PTCHD2	CALHM3	CAV2	ZMAT4	IGDCC3	ADAMTS19	ZCCHC24	USP41	SEMA6A	HSPA8P4	CYR61	WNT11	CAMKV	KLHDC7A	SERPINH1
30	ERO1LB	GNG8	MYB	ITGB5	KIAA1211L	PDZRN4	CWH43	CD34	NELL1	PDZRN4	KLHDC7A	VIM	WNT11	TCP10L	OR5H2	KIF1C	SCN2A	TCP10L	KIAA1024L	GPR17
31	RGS7	GAS2	SLC28A3	TNFRSF10B	RGL3	CAMKV	ADAMTS19	EFEMP2	SCN2A	CLVS1	ALDH3B2	ENG	ST18	PTCHD2	COLCA2	VIM	TMEM178B	CLVS1	ACADSB	FBLN5
32	CNTNAP2	TTC7B	UPK1A	RHOC	RAB3B	PDE1C	ACSS1	CRTAP	RGS7	ATP2B2	VSNL1	CAV1	FOXA2	PDE1C	MOCOS	NOTCH3	NELL1	RBFOX3	SLC12A8	CD93
33	CERS4	TCP10L	MOCOS	CARD6	ERO1LB	GNG8	GNAT3	ZCCHC24	CADPS2	PRIMA1	EN1	IFITM3	SLC36A4	RGS8	ATL2	CAV2	RUNDC3A	GAS2	EFNA4	MYOF
34	DGKB	PDE1C	SLC12A8	EDNRA	SCN2A	PRIMA1	ALDH3B2	CD93	NKX2-1	GNG8	UGT2B28	DLC1	TFF3	HUNK	UGT2B28	SERPINH1	RIMS2	GNG8	IL19	C1QTNF5
35	TFF3	PROKR1	SOSTDC1	REST	WNT11	GAS2	MYB	ENG	DGKB	NPS	FAM124A	DDR2	RAB3B	PLCL2	ALDH3B2	HEPH	ZMAT4	IGFBPL1	LRMP	EDNRA
36	GCH1	EBF3	EFNA4	CAV1	FAM155A	MDGA1	EN1	COL6A2	SPTB	PTCHD2	EFNA4	FBLN5	RGS7	PROKR1	B4GALNT2	CRTAP	USP41	FRMD3	TMPRSS13	CARD6
37	CADPS2	STXBP1	MPV17L	NOTCH2	SLC36A4	FRMD3	COLCA2	CALD1	USP41	FAM71C	ANXA4	PLAT	EPOR	PRIMA1	B4GALT5	RHOC	RGL3	SLC17A7	TULP1	THBS1
38	PRUNE2	MDGA1	BARX2	DAB2	PRUNE2	IGFBPL1	SLC12A8	SERPINH1	CXXC4	KCNQ2	ADSS	DCN	KIAA1211L	GKAP1	EFNA4	MITF	UNC80	TSPAN18	INPP5B	ENG
39	USP41	RBFOX3	TMPRSS13	PPIC	TFF3	RBFOX3	ATL2	DAB2	DBH	PROKR1	MOCOS	IGFBP7	PRUNE2	NTN3	GALNT14	COPZ2	DGKB	MMD2	PLEKHG7	DAB2
40	TOX3	PRDM8	FCHSD2	HMCN1	EPOR	NTNG2	EFNA4	VEGFC	ERO1LB	FRMD3	BARX2	PRSS23	SCN2A	FYN	OR5H6	DAB2	RTBDN	MDGA1	MYB	CTGF
41	LFNG	BVES	KCTD1	MYOF	RNF133	NTN3	BARX2	CTGF	TMEM61	CAMKV	NCOA3	EMILIN1	FAM155A	IGFBPL1	PNOC	PLA2R1	SYT13	EBF3	RGS13	COL12A1
42	SERGEF	NTN3	FOXI1	EPHA2	LFNG	SYT6	OBP2B	FBLN5	RET	PRDM8	C12orf74	PDGFRA	NELL1	NPS	KCTD1	PDGFRA	FOXA2	HUNK	BARX2	CDH11
43	SMOC2	SLC17A7	OBP2B	CNN2	SCIN	ATP2B2	KIAA1024L	PLA2R1	ST18	TSPAN18	ATL2	SERPINH1	CXXC4	PRDM8	RNF223	CARD6	RGS7	RGS8	OBP2B	VEGFC
44	SPTB	PTCHD2	OR5H2	FBLN5	DGKB	HUNK	FAM117A	RHOC	RNF148	RGS8	NREP	CTGF	SMOC2	UMODL1	GNAT3	RBPMS	KIAA1211L	FAM71C	MPV17L	ZCCHC24
45	EXTL3	PGF	ATL2	SCARF1	USP41	NPS	TAF4B	COL12A1	KIAA1244	UMODL1	FCHSD2	CAV2	LFNG	MMD2	OBP2B	CRIM1	ST18	KCNQ2	ACSS1	LDB2
46	SCGN	HUNK	NREP	MRGPRF	ZMAT4	TSPAN18	KCTD1	GNG11	CERS4	HUNK	KCTD1	GPR17	USP20	BARHL1	BARX2	REST	EPOR	CAMK4	C1orf61	COL5A2
47	SCN2A	FAM71C	B4GALT5	EFEMP2	SMOC2	TCP10L	OR5H2	EMILIN1	C1orf127	TTC7B	CCDC115	COL6A3	SERGEF	TSPAN18	CPSF4L	CTGF	SCG2	PRIMA1	FAM124A	LAMA4
48	ST18	IGFBPL1	FAM117A	COL5A2	SERGEF	GKAP1	KAZN	RAB3IL1	RNF133	NTN3	PLEKHG7	LDB2	SH3BP4	FRMD3	RASD2	CCND1	SCGN	BMPER	PIRT	PRSS23
49	EPOR	PRIMA1	ADAMTS19	RAB3IL1	C11orf63	TSHR	C12orf74	CARD6	DNALI1	MMD2	OBP2B	CRIM1	ASXL3	RNF112	TAF4B	CAV1	PRUNE2	TTC7B	SPIC	CRIM1
50	PCSK2	TSPAN18	TFAP2C	AHNAK	CXXC4	BARHL1	CPSF4L	IGFBP7	RAB19	PGF	RNF223	PDGFRB	CADPS2	MDGA1	ADSS	CCDC102A	RAB19	MRAP2	CYP4Z1	VIM

Supplementary Table 3. 80 predictor groups evaluated in the machine learning pipeline (from group 6 to group 10). The purple dashed line separates the 10-gene, 20-gene- 30-gene, 50-gene length signatures.

	G11				G12				G13				G14				G15			
	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1
1	DDC	CERKL	C11orf53	WWTR1	DDC	CERKL	C11orf53	WWTR1	DDC	CERKL	C11orf53	WWTR1	DDC	CERKL	C11orf53	WWTR1	SEC11C	CERKL	C11orf53	OSMR
2	SEC11C	SSTR2	GFI1B	LATS2	SEC11C	SSTR2	TRPM5	LATS2	SEC11C	SSTR2	TRPM5	OSMR	SEC11C	SSTR2	GFI1B	LATS2	DDC	SSTR2	APOBEC1	LATS2
3	STK32A	NHLH1	FAM150A	LAMB2	CNKSR3	NHLH1	GFI1B	MSRB3	CNKSR3	NEUROD4	APOBEC1	LATS2	CNKSR3	NHLH1	PTPN18	MSRB3	CNKSR3	NHLH1	TRPM5	WWTR1
4	PTPRN2	CHRNB4	TRPM5	OSMR	RIMKLA	CHRNB4	APOBEC1	LAMB2	MS4A8	NHLH1	GFI1B	CYBRD1	PTPRN2	CHRNB4	APOBEC1	LAMB2	STK32A	CHRNB4	GFI1B	LAMB2
5	CNKSR3	PPP1R17	APOBEC1	CYBRD1	SCN3A	CHRNA3	FAM150A	OSMR	RGS17	CHRNB4	FAM150A	GPX8	RIMKLA	CHRNA3	IMP4	OSMR	RIMKLA	NEUROD4	FAM150A	MSRB3
6	RIMKLA	NEUROD4	PTPN18	MSRB3	RGS17	LMO1	PTPN18	CYBRD1	RIMKLA	LMO1	BMX	MSRB3	SCN3A	PPP1R17	TRPM5	PTRF	GRP	PPP1R17	PTPN18	CYBRD1
7	MS4A8	CHRNA3	BMX	MYL9	GRP	NEUROD4	IMP4	SYDE1	PTPRN2	CHRNA3	PTPN18	LAMB2	DLL3	NEUROD4	FAM150A	CYBRD1	RGS17	CHRNA3	BMX	EHD2
8	SCN3A	LMO1	ART3	HEPH	MS4A8	PPP1R17	ASCL2	EHD2	SCN3A	PPP1R17	LANCL3	SYDE1	SMPD3	THSD7B	ART3	MYL9	ETS2	LMO1	LANCL3	SYDE1
9	RGS17	THSD7B	LANCL3	SYDE1	STK32A	THSD7B	KLHDC7A	GPX8	STK32A	THSD7B	ART3	EHD2	RGS17	LMO1	VSNL1	HEPH	PTPRN2	SHF	HSPA8P4	LIMS2
10	ETS2	SHF	HES2	PTRF	ETS2	DACH1	LANCL3	PTRF	SMPD3	DACH1	LRMP	LIMS2	STK32A	SHF	HES2	EHD2	SCNN1A	THSD7B	IMP4	PMP22
11	SMPD3	FNDC5	IMP4	GPX8	PTPRN2	SLC17A6	HES2	LIMS2	GRP	PLCL2	CALHM3	PMP22	GRP	SLC17A6	ANO7	SYDE1	NR0B2	CNTN2	ART3	GPX8
12	SCNN1A	DACH1	RGS13	EHD2	NR0B2	SHF	BMX	MRC2	SCNN1A	SHF	IMP4	MYL9	NOL4	NHLH2	PVRL4	LIMS2	SCN3A	NEUROD2	ASCL2	PTRF
13	CA8	SCN1B	LRMP	PMP22	SMPD3	FNDC5	ART3	PMP22	ICA1	SLC17A6	ANO7	MRC2	CNTNAP2	DACH1	FOXI1	GPX8	CAMK1D	HPCA	KLHDC7A	WTIP
14	CNTNAP2	HPCA	CALHM3	LIMS2	CA8	KIAA1614	EN1	MYL9	ETS2	CNTN2	RGS13	WTIP	MS4A8	KIAA1614	COLCA2	PMP22	CA8	FNDC5	CALHM3	MYL9
15	NR0B2	NHLH2	IL19	ZCCHC24	SCNN1A	CNTN2	LRMP	WTIP	NR0B2	FNDC5	VSNL1	IL1R1	CACNA1A	CAMKV	HSPA8P4	AXL	NKX2-1	SCN1B	HES2	AXL
16	GRP	CNTN2	ANO7	AXL	CACNA1A	NEUROD2	C1orf61	AXL	CAMK1D	NEUROD2	ASCL2	PTRF	CA8	NEUROD2	C1orf61	FGD5	ICA1	SLC17A6	LRMP	IL1R1
17	CACNA1A	SLC17A6	SH2D6	COL5A2	RAB3B	HPCA	RGS13	RBMS3	PRUNE2	SCN1B	FOXI1	ITGB5	ICA1	FNDC5	KLHDC7A	COL5A2	SMPD3	PDE1C	SH2D6	RBMS3
18	PRUNE2	IGDCC3	C1orf61	RBPMS	ICA1	GAS2	FAM124A	ITGB5	FOXA2	IGDCC3	IL19	AXL	NR0B2	IGDCC3	FAM124A	MRC2	FBLN7	IGDCC3	RGS13	HEPH
19	ICA1	NEUROD2	HSPA8P4	VIM	PRUNE2	PDZRN4	CALHM3	IL1R1	CA8	PDZRN4	C1orf61	RBMS3	ETS2	SEMA6A	LRMP	WTIP	MS4A8	DACH1	C1orf61	FGD5
20	CADPS2	KIAA1614	VSNL1	WTIP	FOXA2	SCN1B	ANO7	EFEMP2	CACNA1A	NHLH2	HES2	IFITM3	CAMK1D	CNTN2	MYB	ITGB5	ZMAT4	KIAA1614	FAM124A	MRC2
21	RAB3B	GKAP1	ASCL2	IFITM3	NOL4	FAM71C	ALDH3B2	FGD5	RAB19	KIAA1614	ALDH3B2	CAV2	RAB3B	SCN1B	ASCL2	NOTCH3	CACNA1A	CLVS1	ADAMTS19	CAV2
22	WNT11	PLCL2	KLHDC7A	FGD5	NELL1	NHLH2	IL19	HEPH	RGL3	GKAP1	PVRL4	EFEMP2	SCNN1A	HPCA	SH2D6	DDR2	NOL4	PRDM8	MOCOS	CAV1
23	CAMK1D	PRIMA1	FOXI1	IL1R1	DLL3	RGS8	VSNL1	VIM	FBLN7	HPCA	ATL2	CAV1	SLC36A4	PTCHD2	CALHM3	DCN	CNTNAP2	NHLH2	VSNL1	ZCCHC24
24	FBLN7	CLVS1	PVRL4	NOTCH3	SLC36A4	IGDCC3	BARX2	NOTCH3	WNT11	RBFOX3	KLHDC7A	HEPH	PRUNE2	TCP10L	IQCJ	VIM	RAB3B	PLCL2	IL19	IFITM3
25	NOL4	EBF3	UGT2B28	PLA2R1	FBLN7	NPS	SH2D6	COL6A2	LFNG	PRDM8	UGT2B28	GPR17	ST18	PRIMA1	EN1	DLC1	ERO1LB	GNG8	FOXI1	RBPMS
26	FOXA2	MDGA1	MYB	MRC2	CAMK1D	FRMD3	PVRL4	GPR17	NOL4	GAS2	RNF223	RHOC	RIMS2	FRMD3	BMX	IFITM3	SLC36A4	GAS2	ANO7	CRIM1
27	SLC36A4	TSHR	ATL2	RBMS3	RGS7	CAMKV	COLCA2	IFITM3	CADPS2	HUNK	SH2D6	FBLN5	CERS4	GKAP1	RGS13	CRIM1	KIAA1211L	MDGA1	ALDH3B2	NOTCH3
28	SERGEF	ATP2B2	ADAMTS19	ITGB5	CADPS2	TCP10L	CALML5	CAV1	RGS7	PTCHD2	CALML5	NOTCH3	DGKB	GNG8	C12orf74	SERPINH1	NELL1	PDZRN4	COLCA2	ITGB5
29	RNF148	RGS8	RNF223	CRIM1	FAM155A	ATP2B2	MPV17L	RHOC	EPOR	NPS	HSPA8P4	ZCCHC24	SYT4	RGS8	SOSTDC1	ENG	WNT11	GKAP1	MPV17L	PLA2R1
30	RNF133	GNG8	COLCA2	DCN	CXXC4	GNG8	HSPA8P4	ZCCHC24	CERS4	CAMKV	FAM124A	VIM	SERGEF	PLCL2	LANCL3	EMILIN1	TFF3	PRIMA1	EN1	VIM
31	RGS7	SEMA6A	ALDH3B2	CAV1	RNF148	PDE1C	FOXI1	SERPINH1	RAB3B	TCP10L	COLCA2	MYOF	FOXA2	RBFOX3	ALDH3B2	IL1R1	RGS7	MMD2	SLC12A8	RHOC
32	TFF3	TTC7B	FAM124A	EPHA2	SCN2A	MDGA1	MOCOS	ENG	SLC36A4	SEMA6A	MYB	COL5A2	NKX2-1	PDE1C	OBP2B	RBPMS	DGKB	RBFOX3	EFNA4	CALD1
33	ERO1LB	MMP24	ADAT2	EMILIN1	SPTB	RBFOX3	RNF223	COL5A2	KIAA1211L	CAMK4	MPV17L	PLA2R1	RGS7	HUNK	BARX2	CD34	PCSK1	TSPAN18	BARX2	RAB3IL1
34	NKX2-1	RBFOX3	ADSS	SERPINH1	CNTNAP2	SEMA6A	MYB	CAV2	SCN2A	ATP2B2	PRSS21	SERPINH1	WNT11	CLVS1	ADAMTS19	CAV1	SMOC2	CAMKV	PVRL4	EPHA2
35	FAM155A	GAS2	MOCOS	ENG	NKX2-1	TCHH	EFNA4	EDNRA	NKX2-1	TSPAN18	MOCOS	FGD5	NELL1	CAMK4	OR5H2	ZCCHC24	SERGEF	PTCHD2	RNF223	GPR17
36	RAB19	PTCHD2	EN1	CARD6	DBH	GKAP1	OBP2B	RAB3IL1	CNTNAP2	EBF3	BARX2	ENG	SCN2A	GAS2	RNF223	IGFBP7	RGL3	FRMD3	MYB	CTGF
37	NELL1	PROKR1	KCNQ4	FBLN5	KIAA1211L	PRDM8	FCHSD2	EPHA2	TFF3	PROKR1	EN1	DAB2	RUNDC3A	UMODL1	CALML5	COL6A3	CADPS2	RGS8	GNAT3	CYR61
38	INPPL1	CAMKV	CALML5	COL12A1	WNT11	CLVS1	PIRT	CD34	RNF133	MDGA1	ADSS	CTGF	SCGN	IGFBPL1	B4GALT5	EPHA2	CXXC4	IGFBPL1	GNG13	EFEMP2
39	KIAA1211L	PRDM8	GALNT14	MYOF	ZMAT4	PLCL2	ADAMTS19	FBLN5	USP41	RGS8	EFNA4	CYR61	USP41	PDZRN4	MOCOS	CRTAP	CERS4	PGF	AVIL	HMCN1
40	PCSK1	FRMD3	KCTD1	HTRA1	C1orf127	ANKS1B	TFAP2C	CRTAP	RNF148	SLC17A7	CWH43	PRSS23	CADPS2	TSHR	CCDC115	HTRA1	DLL3	SEMA6A	FAM117A	EDNRA
41	ZMAT4	IGFBPL1	KAZN	DDR2	ELOVL7	PTCHD2	B4GALT5	DDR2	SPTB	CLVS1	FCHSD2	EPHA2	FAM155A	EBF3	SLC12A8	CD93	RAB19	TTC7B	UGT2B28	COL5A2
42	CERS4	CAMK4	B4GALT5	FAM198B	SCGN	CORO2B	ACSS1	C1QTNF5	ZMAT4	FAM71C	NCOA3	CPNE8	UNC13A	PROKR1	EFNA4	EDNRA	DBH	TCP10L	GALNT14	CARD6
43	SCN2A	ZBTB18	GNG13	RHOC	RAB19	NTN3	SLC12A8	PLA2R1	ERO1LB	NXPH2	KCNQ4	COL6A2	FBLN7	TTC7B	PLEKHG7	CAV2	GCH1	FYN	OBP2B	HEG1
44	LFNG	PDE1C	EFNA4	PRSS23	HABP2	PRIMA1	TMPRSS13	EMILIN1	DNALI1	IGFBPL1	B4GALT5	THBS1	ERO1LB	NPS	TULP1	RBMS3	GPR98	EBF3	KAZN	CNN1
45	EPOR	PDZRN4	BARX2	CRTAP	PCSK2	FAM212B	KCNQ4	CALD1	CXXC4	RNF112	ADAMTS19	RBPMS	PCSK1	DCC	KCTD1	LDB2	FOXA2	SHD	CALML5	COL6A2
46	PCLO	GRIN2D	SOSTDC1	COL6A2	ENTPD8	HUNK	TAF4B	DLC1	DBH	DENND2A	KCTD1	C1QTNF5	TMEM178B	ATP2B2	SGCG	PLAT	FAM155A	CORO2B	KCNN3	CRTAP
47	CXXC4	CADPS	MPV17L	IGFBP7	USP41	NXPH2	FAM117A	FBN1	SERGEF	ACSL6	ANXA1	COL12A1	GDAP1	FAM212B	ACSS1	ROBO4	RNF148	CADPS	SOSTDC1	TEK
48	GCH1	MMD2	CWH43	CTGF	CERS4	IGFBPL1	ATL2	HTRA1	RET	PRIMA1	PNOC	LAMA4	UNC80	MMD2	ADAT2	SHROOM4	PRUNE2	KLHDC8A	ADSS	CCND1
49	PCSK2	NTN3	PRSS21	CDH11	DNALI1	PROKR1	KCTD1	CRIM1	FAM155A	GNG8	TAF4B	RAB3IL1	RNF148	FAM71C	TAF4B	PDGFRB	SFTA3	RCOR2	KCNQ4	CD93
50	DNALI1	PGF	TSPYL6	CAV2	DGKB	RCOR2	KCNN3	HMCN1	KIAA1244	RCOR2	RASD2	DDR2	SYT1	CORO2B	FAM117A	CARD6	HABP2	ATP2B2	ACSS1	DLC1

Supplementary Table 3. 80 predictor groups evaluated in the machine learning pipeline (from group 11 to group 15). The purple dashed line separates the 10-gene, 20-gene- 30-gene, 50-gene length signatures.

	G16				G17				G18				G19				G20			
	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1	ASCL1	NEUROD1	POU2F3	YAP1
1	SEC11C	CERKL	C11orf53	LATS2	SEC11C	CERKL	C11orf53	WWTR1	SEC11C	CERKL	C11orf53	WWTR1	SEC11C	CERKL	C11orf53	LATS2	DDC	CERKL	C11orf53	LATS2
2	DDC	SSTR2	GFI1B	WWTR1	DDC	SSTR2	GFI1B	LATS2	DDC	SSTR2	TRPM5	LATS2	DDC	SSTR2	APOBEC1	WWTR1	SEC11C	SSTR2	GFI1B	OSMR
3	RIMKLA	CHRNB4	TRPM5	LAMB2	CNKSR3	CHRNB4	APOBEC1	OSMR	CNKSR3	CHRNB4	APOBEC1	OSMR	CNKSR3	NHLH1	TRPM5	LAMB2	RIMKLA	NHLH1	TRPM5	WWTR1
4	CNKSR3	CHRNA3	APOBEC1	OSMR	RGS17	NHLH1	FAM150A	LAMB2	RIMKLA	NHLH1	PTPN18	MSRB3	RIMKLA	NEUROD4	GFI1B	CYBRD1	PTPRN2	NEUROD4	APOBEC1	EHD2
5	SCN3A	NHLH1	FAM150A	CYBRD1	SCN3A	NEUROD4	TRPM5	CYBRD1	PTPRN2	CHRNA3	FAM150A	LAMB2	ETS2	CHRNB4	FAM150A	MSRB3	STK32A	CHRNA3	PTPN18	CYBRD1
6	PTPRN2	NEUROD4	IL19	GPX8	RIMKLA	CHRNA3	PTPN18	MSRB3	SCN3A	LMO1	GFI1B	GPX8	CACNA1A	CHRNA3	ANO7	GPX8	GRP	CHRNB4	FAM150A	LAMB2
7	CACNA1A	LMO1	PTPN18	MSRB3	STK32A	LMO1	ART3	GPX8	RGS17	PPP1R17	IMP4	CYBRD1	SCN3A	PPP1R17	HSPA8P4	OSMR	RGS17	PPP1R17	BMX	MSRB3
8	RGS17	PPP1R17	HES2	ITGB5	PTPRN2	PPP1R17	IMP4	SYDE1	STK32A	NEUROD4	ART3	RBMS3	MS4A8	LMO1	BMX	SYDE1	CNKSR3	LMO1	LANCL3	PTRF
9	STK32A	NEUROD2	CALML5	SYDE1	MS4A8	THSD7B	HES2	HEPH	NOL4	SHF	HES2	SYDE1	SCNN1A	THSD7B	PTPN18	ITGB5	SMPD3	THSD7B	LRMP	LIMS2
10	GRP	CNTN2	IMP4	RBMS3	GRP	SLC17A6	BMX	PTRF	GRP	THSD7B	LANCL3	EHD2	PTPRN2	SHF	LANCL3	PMP22	SCN3A	IGDCC3	KLHDC7A	IL1R1
11	CA8	SLC17A6	BMX	PLA2R1	SMPD3	NEUROD2	KLHDC7A	MYL9	DLL3	CNTN2	FOXI1	WTIP	GRP	NEUROD2	IMP4	MYL9	CNTNAP2	DACH1	IMP4	PMP22
12	NOL4	FNDC5	ART3	WTIP	CA8	FNDC5	LRMP	EHD2	CA8	SLC17A6	CALHM3	MYL9	RGS17	SLC17A6	CALHM3	LIMS2	CAMK1D	CNTN2	HES2	MYL9
13	RAB3B	SHF	LANCL3	EPHA2	NOL4	DACH1	CALHM3	PMP22	CACNA1A	FNDC5	ASCL2	PMP22	SMPD3	DACH1	ART3	WTIP	NKX2-1	SHF	ART3	SYDE1
14	SMPD3	THSD7B	PVRL4	MYL9	NR0B2	SHF	RGS13	WTIP	FBLN7	KIAA1614	BMX	LIMS2	STK32A	FNDC5	ASCL2	MRC2	MS4A8	PLCL2	RGS13	GPX8
15	FBLN7	DACH1	KLHDC7A	PMP22	ETS2	GAS2	C1orf61	NOTCH3	ETS2	NEUROD2	VSNL1	HEPH	CA8	SCN1B	COLCA2	AXL	NR0B2	NEUROD2	C1orf61	IFITM3
16	MS4A8	HPCA	CALHM3	ZCCHC24	FOXA2	SEMA6A	FOXI1	AXL	ICA1	HPCA	KLHDC7A	FGD5	NR0B2	NHLH2	ADAMTS19	EHD2	RGS7	KIAA1614	ASCL2	AXL
17	ETS2	PROKR1	KIAA1024L	EHD2	SCNN1A	SCN1B	FAM124A	LIMS2	MS4A8	NHLH2	PVRL4	PTRF	ICA1	HPCA	LRMP	EPHA2	FAM155A	FNDC5	HSPA8P4	WTIP
18	DLL3	GNG8	ADAMTS19	IFITM3	CAMK1D	HPCA	IL19	MRC2	RAB3B	DACH1	COLCA2	PLA2R1	WNT11	CNTN2	RGS13	PTRF	NOL4	SCN1B	IL19	MRC2
19	SLC36A4	KIAA1614	ASCL2	AXL	ICA1	GNG8	VSNL1	FGD5	SLC36A4	GNG8	MOCOS	CAV2	RAB3B	KIAA1614	VSNL1	NOTCH3	SCNN1A	SLC17A6	CALHM3	ITGB5
20	NR0B2	CLVS1	COLCA2	THBS1	CACNA1A	CNTN2	LANCL3	COL5A2	ZMAT4	GAS2	C1orf61	AXL	CAMK1D	CLVS1	PVRL4	RBMS3	CACNA1A	PDZRN4	SH2D6	CAV1
21	FOXA2	SCN1B	HSPA8P4	MRC2	FBLN7	NHLH2	ASCL2	ITGB5	NR0B2	RBFOX3	CALML5	ZCCHC24	NELL1	GKAP1	ALDH3B2	SERPINH1	ICA1	CLVS1	ALDH3B2	VIM
22	ICA1	GAS2	ALDH3B2	SERPINH1	PRUNE2	IGDCC3	ANO7	PLA2R1	SMPD3	IGDCC3	UGT2B28	MRC2	SLC36A4	PLCL2	HES2	IFITM3	ETS2	GKAP1	VSNL1	CARD6
23	NELL1	NHLH2	C1orf61	HEPH	RAB3B	PRIMA1	PVRL4	IFITM3	ENTPD8	SEMA6A	FAM124A	CAV1	RAB19	EBF3	IL19	IL1R1	ERO1LB	TCP10L	ANO7	CAV2
24	SCNN1A	SEMA6A	ANO7	REST	SERGEF	RBFOX3	HSPA8P4	ZCCHC24	SCNN1A	CAMKV	EN1	CRIM1	FBLN7	PRDM8	SH2D6	PPIC	CA8	HPCA	CALML5	HEPH
25	PRUNE2	EBF3	FOXI1	PTRF	NKX2-1	RGS8	SH2D6	IL1R1	CAMK1D	PROKR1	MYB	IL1R1	CNTNAP2	GAS2	FOXI1	KIF1C	DGKB	RBFOX3	FAM124A	ZCCHC24
26	CAMK1D	RGS8	MOCOS	KIF1C	CNTNAP2	CAMKV	EN1	RBPMS	RIMS2	FAM71C	RNF223	EDNRA	PRUNE2	PDE1C	MOCOS	VIM	RAB3B	PTCHD2	ADAMTS19	RBMS3
27	SCN2A	NTN3	EN1	LIMS2	SLC36A4	PDZRN4	CALML5	RBMS3	PRUNE2	SCN1B	RASD2	RBPMS	FOXA2	ACSL6	CALML5	HEPH	DLL3	NHLH2	FOXI1	RHOC
28	CNTNAP2	PDZRN4	LRMP	NOTCH3	KIAA1211L	TTC7B	UGT2B28	VIM	NELL1	HUNK	OBP2B	EPHA2	CADPS2	IGDCC3	ATL2	PDGFRA	FBLN7	CAMK4	BARX2	FGD5
29	ST18	IGDCC3	VSNL1	CDH11	SCN2A	PRDM8	COLCA2	SERPINH1	NOXA1	RGS8	IL19	EFEMP2	USP41	SEMA6A	PRSS21	FGD5	RGL3	PROKR1	PVRL4	RAB3IL1
30	ZMAT4	PDE1C	UGT2B28	FGD5	CADPS2	CLVS1	MOCOS	HTRA1	SCN2A	PDZRN4	LRMP	DDR2	NOL4	FRMD3	MYB	CCND1	SLC36A4	GNG8	EN1	MYOF
31	WNT11	CAMKV	RGS13	VIM	ERO1LB	PTCHD2	ALDH3B2	EMILIN1	FOXA2	IGFBPL1	ANO7	SERPINH1	RNF148	CAMKV	INPP5B	ZCCHC24	EPOR	CAMKV	EFNA4	TNFRSF10B
32	KCNH6	GKAP1	FAM124A	PDGFRA	WNT11	KIAA1614	ADAMTS19	EPHA2	CNTNAP2	PDE1C	SOSTDC1	FBLN5	KIAA1211L	TSHR	FAM124A	RAB3IL1	FOXA2	PRIMA1	MYB	DAB2
33	ERO1LB	RBFOX3	MYB	IL1R1	ZMAT4	PDE1C	BARX2	DDR2	WNT11	ATP2B2	KCTD1	ITGB5	DLL3	ATP2B2	EN1	REST	CXXC4	MDGA1	COLCA2	TGFBR2
34	DGKB	IGFBPL1	SH2D6	RBPMS	RGS7	PLCL2	MYB	CAV2	CERS4	PTCHD2	EFNA4	IFITM3	NKX2-1	PDZRN4	C1orf61	SHROOM4	JAM3	IGFBPL1	MPV17L	RBPMS
35	NKX2-1	FRMD3	OBP2B	CAV1	DLL3	PGF	EFNA4	CRIM1	RGL3	PRDM8	SH2D6	THBS1	DGKB	IGFBPL1	NCOA3	COL5A2	UNC80	EBF3	MOCOS	CRTAP
36	EXTL3	ATP2B2	FAM117A	CARD6	USP41	ATP2B2	FAM117A	CAV1	SERGEF	GKAP1	ALDH3B2	NOTCH3	CERS4	NTNG2	C12orf74	CYR61	PRUNE2	PDE1C	AVIL	IL4R
37	USP41	TCP10L	EFNA4	CRTAP	DGKB	FRMD3	SLC12A8	MYOF	UNC80	TCP10L	OR5H6	C1QTNF5	RGS7	RGS8	FAM117A	CAV1	ZMAT4	RGS8	FAM117A	COL6A2
38	RIMS2	MMP24	ATL2	CAV2	DNALI1	GKAP1	ATL2	COL12A1	EXTL3	MRAP2	OR5H2	HTRA1	ERO1LB	TCP10L	ADSS	PRSS23	PCSK1	HUNK	ADAT2	PLA2R1
39	RGS7	PTCHD2	SOSTDC1	FAM198B	CERS4	NPS	SOSTDC1	RHOC	SCG2	FRMD3	CWH43	VIM	TFF3	NTN3	ACSS1	EDNRA	SYT4	CORO2B	B4GALT5	EPHA2
40	GNAO1	ASIC1	B4GALT5	COL12A1	NELL1	MDGA1	PIRT	CARD6	C1orf127	EBF3	TAF4B	ENG	RNF133	UMODL1	MPV17L	GPR17	KIAA1244	FYN	NCOA3	COL5A2
41	CADPS2	PRIMA1	FCHSD2	RAB3IL1	RNF148	PROKR1	MPV17L	PLAT	GDAP1	RCOR2	CCDC115	HMCN1	USP20	FAM71C	SLC28A3	HMCN1	SPTB	MMD2	FCHSD2	AHNAK
42	UNC80	STXBP1	KCTD1	RHOC	TFF3	FAM71C	ADTRP	ENG	EPOR	CHRNA5	HSPA8P4	MRGPRF	INPPL1	MDGA1	RNF223	FBLN5	CADPS2	NRXN1	KCTD1	NNMT
43	SERGEF	RCOR2	KAZN	EDNRA	RNF133	NTN3	OBP2B	FBLN5	DBH	MMD2	ACSS1	PRKG1	BEST3	NPS	GNAT3	RHOC	WNT11	RCOR2	RNF223	GPR17
44	PCSK2	NPS	BARX2	PPIC	EPOR	EBF3	B4GALT5	COL6A3	TFF3	ANKS1B	TULP1	MYOF	CXXC4	PROKR1	KLHDC7A	CTGF	CERS4	STXBP1	UGT2B28	CYR61
45	ENTPD8	PRDM8	TMPRSS13	PLD1	PCSK2	TCHH	FCHSD2	COPZ2	USP41	CLVS1	BARX2	RHOC	SCN2A	PRIMA1	SLC12A8	PLA2R1	NELL1	SEMA6A	C12orf74	CALD1
46	SPTB	FAM71C	NREP	CCDC102A	ST18	TSPAN18	IQCJ	CRTAP	RUNDC3A	MDGA1	B4GALT5	CD93	RGL3	PTCHD2	ADTRP	CAV2	SFTA3	UMODL1	KAZN	ENG
47	SNCAIP	TTC7B	MPV17L	CYR61	FAM155A	MMD2	KAZN	DCN	NKX2-1	NTN3	ACADSB	CTGF	DBH	MRAP2	KAZN	RBPMS	RAB19	FRMD3	GALNT14	THBS1
48	UNC13A	MDGA1	ADTRP	HTRA1	TMEM61	HUNK	NREP	PRSS23	SMOC2	TSPAN18	ADAMTS19	PLAT	SERGEF	TCHH	B4GALT5	CD34	GCH1	TSPAN18	INPP5B	NOTCH3
49	SMOC2	CORO2B	OR5H2	MRGPRF	SPTB	TCP10L	TMPRSS13	EFEMP2	KSR2	PRIMA1	SLC12A8	SHROOM4	DNALI1	CAMK4	UGT2B28	F2R	SERGEF	SBSPON	ATL2	EMILIN1
50	EPOR	PCDH8	ANXA4	MYOF	PCSK1	RNF112	KCNQ4	IGFBP7	ST18	PLCL2	RGS13	DLC1	DPYSL3	DENND2A	ADAT2	COL6A2	PCLO	KLHDC8A	ADSS	SCARF1

Supplementary Table 3. 80 predictor groups evaluated in the machine learning pipeline (from group 16 to group 20). The purple dashed line separates the 10-gene, 20-gene- 30-gene, 50-gene length signatures. Highlighted cells in purple represent the 20-gene subtype-specific downstream programs used as features in our final NAPY SVM classifier.

10. References

1. **Rudin, C. M., Brambilla, E., Faivre-Finn, C., Sage, J.** *Small-cell lung cancer.* s.l. : Nature Reviews Disease Primers, 2021. Vol. 7.

2. **Rudin, C. M., Poirier, J. T., Byers, L. A., Dive, C., Dowlati, A., George, J., Heymach, J. v., Johnson, J. E., Lehman, J. M., MacPherson, D., Massion, P. P., Minna, J. D., Oliver, T. G., Quaranta, V., Sage, J., Thomas, R. K., Vakoc, C. R., & Gazdar, A.** *Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data.* s.l. : Nature Reviews Cancer, 2019. Vol. 19.

3. **Nicholson, A. G., Tsao, M. S., Beasley, M. B., Borczuk, A. C., Brambilla, E., Cooper, W. A., Dacic, S., Jain, D., Kerr, K. M., Lantuejoul, S., Noguchi, M., Papotti, M., Rekhtman, N., Scagliotti, G., van Schil, P., Sholl, L., Yatabe, Y., Yoshida, A., &.** *The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015.* s.l. : Journal of Thoracic Oncology, 2022. Vol. 17.

4. **Szczepanski, A. P., Tsuboyama, N., Watanabe, J., Hashizume, R., Zhao, Z., & Wang, L.** *POU2AF2/C11orf53 functions as a coactivator of POU2F3 by maintaining chromatin accessibility and enhancer activity.* s.l. : Science Advances, 2022. Vol. 8.

5. **Chan, J. M., Quintanal-Villalonga, Á., Gao, V. R., Xie, Y., Allaj, V., Chaudhary, O., Masilionis, I., Egger, J., Chow, A., Walle, T., Mattar, M., Yarlagadda, D. V. K., Wang, J. L., Uddin, F., Offin, M., Ciampricotti, M., Qeriqi, B., Bahr, A., de Stanchina.** *Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer.* s.l. : Cancer Cell, 2021. pp. 1479-1496. Vol. 39.

6. **Baine, M. K., Hsieh, M. S., Lai, W. V., Egger, J. v., Jungbluth, A. A., Daneshbod, Y., Beras, A., Spencer, R., Lopardo, J., Bodd, F., Montecalvo, J., Sauter, J. L., Chang, J. C., Buonocore, D. J., Travis, W. D., Sen, T., Poirier, J. T., Rudin, C. M., & Re.** *SCLC Subtypes Defined by ASCL1, NEUROD1, POU2F3, and YAP1: A Comprehensive Immunohistochemical and Histopathologic Characterization.* s.l. : Journal of Thoracic Oncology, 2020. pp. 1823–1835. Vol. 15.

7. **Andersson-Rolf, A., Clevers, H., & Dayton, T. L.** *Diffuse Hormonal Systems - Endotext -.* s.l. : NCBI Bookshelf, 2021.

8. **Metovic, J., Barella, M., & Pelosi, G.** *Neuroendocrine neoplasms of the lung: a pathology update.* s.l. : Memo - Magazine of European Medical Oncology, 2021. Vol. 14.

9. **Zhang, W., Girard, L., Zhang, Y. A., Haruki, T., Papari-Zareei, M., Stastny, V., Ghayee, H. K., Pacak, K., Oliver, T. G., Minna, J. D., & Gazdar, A. F.** *Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes.* s.l. : Translational Lung Cancer Research,, 2018. Vol. 7.

10. **Ireland, A. S., Micinski, A. M., Kastner, D. W., Guo, B., Wait, S. J., Spainhower, K. B., Conley, C. C., Chen, O. S., Guthrie, M. R., Soltero, D., Qiao, Y., Huang, X., Tarapcsák, S., Devarakonda, S., Chalishazar, M. D., Gertz, J., Moser, J. C., Marth, G.,.** *MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate.* s.l. : Cancer Cell, 2020. Vol. 38.

11. **Cao, S., Wang, Y., Zhou, Y., Zhang, Y., Ling, X., Zhang, L., Li, J., Yang, Y., Wang, W., Shurin, M. R., & Zhong, H.** *A Novel Therapeutic Target for Small-Cell Lung Cancer: Tumor-Associated Repair-like Schwann Cells.* s.l. : Cancers, 2022. Vol. 14.

12. **Ding, X. L., Su, Y. G., Yu, L., Bai, Z. L., Bai, X. H., Chen, X. Z., Yang, X., Zhao, R., He, J. X., & Wang, Y. Y.** *Clinical characteristics and patient outcomes of molecular subtypes of small cell lung cancer (SCLC).* s.l. : World Journal of Surgical Oncology, 2022. Vol. 20.

13. **Qu, S., Fetsch, P., Thomas, A., Pommier, Y., Schrump, D. S., Miettinen, M. M., & Chen, H.** *Molecular Subtypes of Primary SCLC Tumors and Their Associations With Neuroendocrine and Therapeutic Markers.* s.l. : Journal of Thoracic Oncology, 2022. pp. 141–153. Vol. 17.

14. **Qi, J., Zhang, J., Liu, N., Zhao, L., & Xu, B.** *Prognostic Implications of Molecular Subtypes in Primary Small Cell Lung Cancer and Their Correlation With Cancer Immunity.* s.l. : Frontiers in Oncology, 2022. Vol. 12.

15. **Gay, C. M., Stewart, C. A., Park, E. M., Diao, L., Groves, S. M., Heeke, S., Nabet, B. Y., Fujimoto, J., Solis, L. M., Lu, W., Xi, Y., Cardnell, R. J., Wang, Q., Fabbri, G., Cargill, K. R., Vokes, N. I., Ramkumar, K., Zhang, B., della Corte, C. M., … Byer.** *Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities.* s.l. : Cancer Cell, 2021. Vol. 39.

16. **Tian, Y., Li, Q., Yang, Z., Zhang, S., Xu, J., Wang, Z., Bai, H., Duan, J., Zheng, B., Li, W., Cui, Y., Wang, X., Wan, R., Fei, K., Zhong, J., Gao, S., He, J., Gay, C. M., Zhang, J., … Tang, F.** *Single-cell transcriptomic profiling reveals the tumor heterogeneity of small-cell lung cancer.* s.l. : Signal Transduction and Targeted Therapy, 2022. Vol. 7.

17. **Schwendenwein, A., Megyesfalvi, Z., Barany, N., Valko, Z., Bugyik, E., Lang, C., Ferencz, B., Paku, S., Lantos, A., Fillinger, J., Rezeli, M., Marko-Varga, G., Bogos, K., Galffy, G., Renyi-Vamos, F., Hoda, M. A., Klepetko, W., Hoetzenecker, K., Laszlo, V.** *Molecular profiles of small cell lung cancer subtypes: therapeutic implications. .* s.l. : Molecular Therapy - Oncolytics, 2021. pp. 470–483. Vol. 20.

18. **Groves, S. M., Ildefonso, G. v., McAtee, C. O., Ozawa, P. M. M., Ireland, A. S., Stauffer, P. E., Wasdin, P. T., Huang, X., Qiao, Y., Lim, J. S., Bader, J., Liu, Q., Simmons, A. J., Lau, K. S., Iams, W. T., Hardin, D. P., Saff, E. B., Holmes, W. R., Tyson.** *Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancerCell Systems, 13(9). https://doi.org/10.1016/j.cels.2022.07.006.* s.l. : Cell Systems, 2022. Vol. 13.

19. **Poirier, J. T., George, J., Owonikoko, T. K., Berns, A., Brambilla, E., Byers, L. A., Carbone, D., Chen, H. J., Christensen, C. L., Dive, C., Farago, A. F., Govindan, R., Hann, C., Hellmann, M. D., Horn, L., Johnson, J. E., Ju, Y. S., Kang, S., Krasnow, M.** *New Approaches to SCLC Therapy: From the Laboratory to the Clinic.* s.l. : Journal of Thoracic Oncology, 2020. Vol. 15.

20. **Keogh, A., Finn, S., & Radonic, T.** *Emerging Biomarkers and the Changing Landscape of Small Cell Lung Cancer.* s.l. : Cancers, 2022. Vol. 14.

21. **Kukurba, K. R., & Montgomery, S. B.** *RNA sequencing and analysis.* s.l. : Cold Spring Harbor Protocols, 2015. pp. 951–969. Vol. 2015.

22. **Kreis, J., Nedić, B., Mazur, J., Urban, M., Schelhorn, S. E., Grombacher, T., Geist, F., Brors, B., Zühlsdorf, M., & Staub, E.** *RosettaSX: Reliable gene expression signature scoring of cancer models and patients.* United States : Neoplasia , 2021. Vol. 23.

23. **James, G. et al.** *An introduction to statistical learning: With applications in R.* Boston : Springer, 2021.

24. **Karatzoglou, A. and Smola , A.** *Kernlab: Kernel-based Machine Learning Lab, Package 'kernlab'.* 2023.

25. **Bawono, A. H., Abdurrahman Bachtiar, F., Supianto, A. A., Komputer, F. I., & Brawijaya, U.** *Nearest Centroid Classifier with Centroid-Based Outlier Removal for Classification.* s.l. : Journal of Information Technology and Computer Science, 2020. Vol. 5.

26. **Parvandeh, S., Yeh, H. W., Paulus, M. P., & McKinney, B. A.** *Consensus features nested cross-validation.* s.l. : Bioinformatics, 2020. Vol. 36.

27. **Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. v., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. v., Pa.** *Next-generation characterization of the Cancer Cell Line Encyclopedia.* s.l. : Nature, 2019. Vol. 569.

28. **Rudin, C. M., Balli, D., Lai, W. V., Richards, A. L., Nguyen, E., Egger, J. v., Choudhury, N. J., Sen, T., Chow, A., Poirier, J. T., Geese, W. J., Hellmann, M. D., & Forslund, A.** *Clinical Benefit From Immunotherapy in Patients With SCLC Is Associated With Tumor Capacity for Antigen Presentation.* s.l. : Journal of Thoracic Oncology, 2023.

29. **Aybey, B., Zhao, S., Brors, B., & Staub, E.** *Immune cell type signature discovery and random forest classification for analysis of single cell gene expression datasets.* s.l. : Frontiers in Immunology, 2023. Vol. 14.

30. **Braune, E. B., Geist, F., Tang, X., Kalari, K., Boughey, J., Wang, L., Leon-Ferre, R. A., D'Assoro, A. B., Ingle, J. N., Goetz, M. P., Kreis, J., Wang, K., Foukakis, T., Seshire, A., Wienke, D., & Lendahl, U.** *Identification of a Notch transcriptomic signature for breast cancer.* s.l. : Breast Cancer Research, 2024. Vol. 26.

31. **Kishore, J., Goel, M., & Khanna, P.** *Understanding survival analysis: Kaplan-Meier estimate.* s.l. : International Journal of Ayurveda Research, 2010. Vol. 1.

32. **Pozo, K., Kollipara, R. K., Kelenis, D. P., Rodarte, K. E., Ullrich, M. S., Zhang, X., Minna, J. D., & Johnson, J. E.** *ASCL1, NKX2-1, and PROX1 co-regulate subtype-specific genes in small-cell lung cancer.* s.l. : IScience, 2021. Vol. 24.

33. **Borromeo, M. D., Savage, T. K., Kollipara, R. K., He, M., Augustyn, A., Osborne, J. K., Girard, L., Minna, J. D., Gazdar, A. F., Cobb, M. H., & Johnson, J. E.** *ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs.* s.l. : Cell Reports, 2016. Vol. 16.

34. **Huang, Y. H., Klingbeil, O., He, X. Y., Wu, X. S., Arun, G., Lu, B., Somerville, T. D. D., Milazzo, J. P., Wilkinson, J. E., Demerdash, O. E., Spector, D. L., Egeblad, M., Shi, J., & Vakoc, C. R.** *POU2F3 is a master regulator of a tuft cell-like variant of small lung cancer.* s.l. : Genes and Development, 2018. pp. 13-14. Vol. 32.

35. **Kudoh, S., Tenjin, Y., Kameyama, H., Ichimura, T., Yamada, T., Matsuo, A., Kudo, N., Sato, Y., & Ito, T.** *Significance of achaete-scute complex homologue 1 (ASCL1) in pulmonary neuroendocrine carcinomas; RNA sequence analyses using small cell lung cancer cells and Ascl1-induced pulmonary neuroendocrine carcinoma cells.* s.l. : Histochemistry and Cell Biology, 2020. Vol. 153.

36. **Sivakumar, S., Moore, J. A., Montesion, M., Sharaf, R., Lin, D. I., Colón, C. I., Fleishmann, Z., Ebot, E. M., Newberg, J. Y., Mills, J. M., Hegde, P. S., Pan, Q., Dowlati, A., Frampton, G. M., Sage, J., & Lovly, C. M.** *Integrative Analysis of a Large Real-World Cohort of Small Cell Lung Cancer Identifies Distinct Genetic Subtypes and Insights into Histologic Transformation.* s.l. : Cancer Discovery, 2023. pp. 1572–1591. Vol. 13.

37. **Kashima, J., & Okuma, Y.** *Advances in biology and novel treatments of SCLC: The four-color problem in uncharted territory. In Seminars in Cancer Biology.* s.l. : Academic Press, 2022. pp. 386–395. Vol. 86.

38. **Tlemsani, C., Pongor, L., Elloumi, F., Girard, L., Huffman, K. E., Roper, N., Varma, S., Luna, A., Rajapakse, V. N., Sebastian, R., Kohn, K. W., Krushkal, J., Aladjem, M. I., Teicher, B. A., Meltzer, P. S., Reinhold, W. C., Minna, J. D., Thomas, A., & Pom.** *SCLC-CellMiner: A Resource for Small Cell Lung Cancer Cell Line Genomics and Pharmacology Based on Genomic Signatures.* s.l. : Cell Reports, 2020. Vol. 33.

39. **Kawai, H., Matsuoka, R., Ito, T., & Matsubara, D.** *Molecular Subtypes of High-Grade Neuroendocrine Carcinoma (HGNEC): What is YAP1-Positive HGNEC?* s.l. : Frontiers in Bioscience, 2022. Vol. 27.

40. **Pearsall, S. M., Humphrey, S., Revill, M., Morgan, D., Frese, K. K., Galvin, M., Kerr, A., Carter, M., Priest, L., Blackhall, F., Simpson, K. L., & Dive, C.** *The Rare YAP1 Subtype of SCLC Revisited in a Biobank of 39 Circulating Tumor Cell Patient Derived Explant Models: A Brief Report.* s.l. : Journal of Thoracic Oncology, 2020. Vol. 15.

41. **Pearsall, S. M., Williamson, S. C., Humphrey, S., Hughes, E., Morgan, D., García Marqués, F. J., Awanis, G., Carroll, R., Burks, L., Shue, Y. T., Bermudez, A., Frese, K. K., Galvin, M., Carter, M., Priest, L., Kerr, A., Zhou, C., Oliver, T. G., Humphries,.** *Lineage plasticity in SCLC generates non-neuroendocrine cells primed for vasculogenic mimicry .* s.l. : Journal of Thoracic Oncology, 2023.

42. **Simpson, K. L., Stoney, R., Frese, K. K., Simms, N., Rowe, W., Pearce, S. P., Humphrey, S., Booth, L., Morgan, D., Dynowski, M., Trapani, F., Catozzi, A., Revill, M., Helps, T., Galvin, M., Girard, L., Nonaka, D., Carter, L., Krebs, M. G., … Dive, C.** *A biobank of small cell lung cancer CDX models elucidates inter- and intratumoral phenotypic heterogeneity.* s.l. : Nature Cancer, 2020. Vol. 1.

43. **Megyesfalvi, Z., Barany, N., Lantos, A., Valko, Z., Pipek, O., Lang, C., Schwendenwein, A., Oberndorfer, F., Paku, S., Ferencz, B., Dezso, K., Fillinger, J., Lohinai, Z., Moldvay, J., Galffy, G., Szeitz, B., Rezeli, M., Rivard, C., Hirsch, F. R., … Dome,.** *Expression patterns and prognostic relevance of subtype-specific transcription factors in surgically resected small-cell lung cancer: an international multicenter study.* s.l. : Journal of Pathology, 2022. pp. 674–686. Vol. 257.

44. **Baine, M. K., Febres-Aldana, C. A., Chang, J. C., Jungbluth, A. A., Sethi, S., Antonescu, C. R., Travis, W. D., Hsieh, M. S., Roh, M. S., Homer, R. J., Ladanyi, M., Egger, J. v., Lai, W. V., Rudin, C. M., & Rekhtman, N.** *POU2F3 in SCLC: Clinicopathologic and Genomic Analysis With a Focus on Its Diagnostic Utility in Neuroendocrine-Low SCLC.* s.l. : Journal of Thoracic Oncology, 2022. Vol. 17.

45. **George, J., Lim, J. S., Jang, S. J., Cun, Y., Ozretia, L., Kong, G., Leenders, F., Lu, X., Fernández-Cuesta, L., Bosco, G., Müller, C., Dahmen, I., Jahchan, N. S., Park, K. S., Yang, D.,**

Karnezis, A. N., Vaka, D., Torres, A., Wang, M. S., … Thomas, R. *Comprehensive genomic profiles of small cell lung cancer.* s.l. : Nature, 2015. Vol. 524.

46. **Jiao, S., Zhang, X., Wang, D., Fu, H., & Xia, Q.** *Genetic Alteration and Their Significance on Clinical Events in Small Cell Lung Cancer.* s.l. : Cancer Management and Research, 2022. pp. 1493-1505. Vol. 14.

47. **Hu, J., Wang, Y., Zhang, Y., Yu, Y., Chen, H., Liu, K., Yao, M., Wang, K., Gu, W., & Shou, T.** *Comprehensive genomic profiling of small cell lung cancer in Chinese patients and the implications for therapeutic potential.* s.l. : Cancer Medicine, 2019. pp. 4338–4347. Vol. 8.

48. **W Tan, W., & Maghfoor, I.** *Small cell lung cancer (SCLC) treatment & management.* 2023.

49. **Chen, P., Sun, C., Wang, H., Zhao, W., Wu, Y., Guo, H., Zhou, C., & He, Y.** *YAP1 expression is associated with survival and immunosuppression in small cell lung cancer.* s.l. : Cell Death and Disease, 2023. Vol. 14.

50. **Li, N., Lu, N., & Xie, C.** *The Hippo and Wnt signalling pathways: crosstalk during neoplastic progression in gastrointestinal tissue.* s.l. : FEBS Journal, 2019. Vol. 286.

51. **Varelas, X., Miller, B. W., Sopko, R., Song, S., Gregorieff, A., Fellouse, F. A., Sakuma, R., Pawson, T., Hunziker, W., McNeill, H., Wrana, J. L., & Attisano, L.** *he Hippo Pathway Regulates Wnt/β-Catenin Signaling.* s.l. : Developmental Cell, 2010. Vol. 18.

52. **Imajo, M., Miyatake, K., Iimura, A., Miyamoto, A., & Nishida, E.** *A molecular mechanism that links Hippo signalling to the inhibition of Wnt/β-catenin signalling.* s.l. : EMBO Journal,, 2012. Vol. 31.

53. **Yatabe, Y.** *Reassessing the SCLC Subtypes.* s.l. : Journal of Thoracic Oncology, 2020. Vol. 15.

54. **Brägelmann, J., Böhm, S., Guthrie, M. R., Mollaoglu, G., Oliver, T. G., & Sos, M. L.** *Family matters: How MYC family oncogenes impact small cell lung cancer.* s.l. : Cell Cycle, 2017. Vol. 16.

55. **Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., … Haye.** *Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.* s.l. : Cancer Cell, 2010. Vol. 17.

56. **Kang, J., D'Andrea, A. D., & Kozono, D.** *A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy.* s.l. : Journal of the National Cancer Institute, 2012. Vol. 104.

57. **Bennett, L., Palucka, A. K., Arce, E., Cantrell, V., Borvak, J., Banchereau, J., & Pascual, V.** *Interferon and granulopoiesis signatures in systemic lupus erythematosus blood.* s.l. : Journal of Experimental Medicine, 2003. Vol. 197.

58. **Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., Bruneval, P., Fridman, W. H., Becker, C., Pagès, F., Speicher, M. R., Trajanoski, Z., & Galon, J. Ô.** *Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer.* s.l. : Immunity, 2013. Vol. 39.

59. **Mischel, P. S., Shai, R., Shi, T., Horvath, S., Lu, K. v., Choe, G., Seligson, D., Kremen, T. J., Palotie, A., Liau, L. M., Cloughesy, T. F., & Nelson, S. F.** *Identification of molecular subtypes of glioblastoma by gene expression profiling.* s.l. : Oncogene, 2003. Vol. 22.

60. **Zhang, J., Smolen, G. A., & Haber, D. A.** *Negative regulation of YAP by LATS1 underscores evolutionary conservation of the Drosophila Hippo pathway.* s.l. : Cancer Research, 2008. Vol. 68.

61. **Heinonen, H., Nieminen, A., Saarela, M., Kallioniemi, A., Klefström, J., Hautaniemi, S., & Monni, O.** *Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer.* s.l. : BMC Genomics, 2008. Vol. 9.

62. **Mazur, J., Kreis, J., Trivier, E., Dillon, C., Wienke, D., & Staub, E.** *Abstract A38: A 4-gene YAP-related pathway expression signature informs about dependence of tumors on Hippo pathway signaling.* s.l. : Molecular Cancer Research, 2020. Vol. 18.

63. **Verrecchia, F., Chu, M. L., & Mauviel, A.** *Identification of Novel TGF-β/Smad Gene Targets in Dermal Fibroblasts using a Combined cDNA Microarray/Promoter Transactivation Approach.* s.l. : Journal of Biological Chemistry, 2001. Vol. 276.

64. **Taube, J. H., Herschkowitz, J. I., Komurov, K., Zhou, A. Y., Gupta, S., Yang, J., Hartwell, K., Onder, T. T., Gupta, P. B., Evans, K. W., Hollier, B. G., Ram, P. T., Lander, E. S., Rosen, J. M., Weinberg, R. A., & Mani, S. A.** *Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes.* s.l. : Proceedings of the National Academy of Sciences of the United States of America, 2010. Vol. 107.

65. **Kool, M., Koster, J., Bunt, J., Hasselt, N. E., Lakeman, A., van Sluis, P., Troost, D., Schouten-van Meeteren, N., Caron, H. N., Cloos, J., Mršić, A., Ylstra, B., Grajkowska, W., Hartmann, W., Pietsch, T., Ellison, D., Clifford, S. C., & Versteeg, R.** *Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features.* s.l. : PLoS ONE, 2008. Vol. 3.

66. **Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P.** *The Molecular Signatures Database Hallmark Gene Set Collection.* s.l. : Cell Systems, 2015. Vol. 1.

11. Acknowledgements

First, I would like to thank my family and friends, for their constant support throughout these years.

A special thanks to my supervisors, Eike and Lu, who have guided me during the whole process, for their patience, motivation, and exemplary standards that inspired me to give my very best.

A special mention to Julian, a colleague from the Bioinformatics team at Merck, for the valuable discussions, support, and guidance.

Finally, I extend my appreciation to the team at Merck for their openness and flexibility, which were instrumental in the successful completion of my thesis.