Integrating somatic and germline multi-omics data to improve our understanding of lung cancer : a computational biology perspective

Aurélie Gabriel

N°d'ordre NNT : 2020LYSE1324

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED340
BIOLOGIE MOLÉCULAIRE INTÉGRATIVE ET CELLULAIRE (BMIC)

Spécialité de doctorat : Oncogénomique

Soutenue publiquement le 09/12/2020, par :
Aurélie GABRIEL

Integrating somatic and germline multi-omics data to improve our understanding of lung cancer: a computational biology perspective.

Devant le jury composé de :

DUMONTET, Charles	Professeur (PU-PH)	Université Lyon1 UMR 5286 - CRCL	Président
RELTON, Caroline	Professeur	University of Bristol	Rapporteure
THIRLWELL, Chrissie	Professeur	University of Exeter	Rapporteure
AMOS, Christopher	Professeur	Baylor College of Medicine	Examinateur
DUMONTET, Charles	Professeur (PU-PH)	Université Lyon1 UMR 5286 - CRCL	Examinateur
MCKAY, James	Chercheur	CIRC	Directeur de thèse
FOLL, Matthieu	Chercheur	CIRC	Co-directeur de thèse

Université Claude Bernard – LYON 1

Administrateur provisoire de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Jean-François MORNEX
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

UNIVERSITÉ CLAUDE BERNARD LYON 1

Abstract

Doctor of Philosophy

Integrating somatic and germline multi-omics data to improve our understanding of lung cancer: a computational biology perspective.

by Aurélie GABRIEL

Cancer is a complex disease caused by endogenous and exogenous factors and impacting multiple omics layers. In the past decades, the high-resolution interrogation of these layers provided valuable insights on cancer etiology and development. In the case of lung cancer, both germline susceptibility and somatic landscapes were widely explored. However, the identification of disease-causal pathways is still challenging and certain cancer types remain understudied. Molecular characterization of lung cancer could thus provide new insights. Such characterization requires though to apply computational methods adapted to the complexity and the high dimensionality of omics data. In this thesis, we took advantage of integrative analyses to explore lung cancer omics data. Firstly, multi-omics data were integrated for the molecular characterization of lung neuroendocrine neoplasms. Machine learning methods identified molecular subgroups which had distinct prognosis and were clinically relevant. Subsequently, a molecular map integrating six previously published transcriptomic datasets was built. The map corroborated previous biological hypotheses and was designed to encourage the generation of new hypotheses by providing the underlying homogenized dataset as well as resources promoting reproducibility and data reuse. Finally, the interplay between germline and somatic layers in lung cancers have been explored. Associations between germline susceptibility to lung cancer and mutational burden in lung tumors were identified. While tobacco smoking susceptibility SNPs were major drivers, pleiotropic effects were also detected, suggesting that other pathways might be involved.

Keywords: Omics data, Computational biology, Data integration, Lung cancer

Intégration de données multi-omiques constitutionnelles et somatiques, pour une meilleure compréhension du cancer du poumon : une approche computationnelle.

Résumé

Le cancer est une maladie complexe causée par des facteurs endogènes et exogènes, impactant différentes couches omiques. Au cours des dernières décennies, l'interrogation de ces couches à haute résolution a permis d'étudier l'étiologie et le développement des cancers. Dans le cas du cancer du poumon, les profiles constitutionnels et somatiques ont été largement explorés. Toutefois, l'identification des gènes responsables de la maladie reste limitée et certains types de cancer sont peu étudiés. La caractérisation moléculaire des cancers du poumon pourrait donc améliorer les connaissances actuelles. Elle nécessite cependant l'application de méthodes adaptées à la complexité et la grande dimension des données omiques. Dans cette thèse, nous avons mené plusieurs approches intégratives pour explorer ces données au sein du cancer du poumon. Premièrement, nous avons intégré des données multi-omiques pour décrire les tumeurs neuroendocrines (TNE) du poumon et avons identifié, grâce à des méthodes d'apprentissage automatique, des groupes moléculaires de pronostic différents. Par la suite, une carte moléculaire intégrant six jeux de données transcriptomiques de TNE du poumon a été établie afin de favoriser la génération de nouvelles hypothèses et la réutilisation des données. Enfin, nous avons exploré l'interaction entre les évènements constitutionnels et somatiques au sein des cancers du poumon. Une association, majoritairement due à la susceptibilité au tabac, a été détectée entre les variants constitutionnels et la charge mutationnelle des tumeurs. Cependant, des effets pléiotropiques ont également été détectés, suggérant que d'autres mécanismes pourraient être impliqués.

Mots clés: Données omiques, Biologie computationnelle, Intégration de données, Cancers du poumon.

Intégration de données multi-omiques constitutionnelles et somatiques, pour une meilleure compréhension du cancer du poumon : une approche computationnelle.

Résumé en français

Le cancer est une des premières causes de décès à travers le monde. La maladie peut toucher tous les organes du corps humain et tient son origine d'une cellule normale dont le génome, porteur de notre matériel génétique, a été altéré. Les altérations génomiques s'accumulent au fil du temps et confèrent aux cellules cancéreuses certaines caractéristiques biologiques permettant une prolifération non contrôlée. Ces altérations peuvent correspondre à des mutations génétiques mais aussi à des modifications épigénétiques comme la méthylation de l'ADN ou les modifications d'histones. Elles influencent donc diverses étapes nécessaires à l'expression de gènes en protéines et diverses couches omiques telles que le génome, le transcriptome, le méthylome ou encore le protéome sont donc impactées. En ce qui concerne les mutations, deux catégories existent. La première catégorie de mutations est présente dans les cellules germinales et est héritée de nos parents à la naissance. Ces mutations, dites constitutionnelles, sont donc détectables dans toutes les cellules de notre corps, aussi bien dans les cellules normales qu'au sein des cellules cancéreuses chez un individu atteint de cancer. La seconde catégorie de mutations est acquise tout au long de notre vie, on parle de mutations somatiques. Elles peuvent être le résultat de mécanismes endogènes, comme l'apparition d'erreurs lors de la réplication de l'ADN, mais aussi de phénomènes exogènes, par exemple à la suite d'expositions environnementales telles que le tabac ou encore les rayons UV. Chacun de ces évènements laisse des traces qui peuvent être détectées en analysant les différentes couches omiques. L'accumulation de mutations, par exemple, génère ce qu'on appelle des signatures mutationnelles, propres à chaque processus mutationnel. Aussi, des profils d'expression ou de méthylation peuvent refléter diverses activités cellulaires propres à chaque type de tumeurs et permettre de les distinguer.

Au cours des dernières décennies, les technologies développées en génomique et épigénétique ont permis d'interroger les données omiques à haute résolution afin d'étudier l'étiologie et le développement des cancers, tout d'abord à petite échelle puis par le biais de projets de grande échelle favorisant le partage de données au sein de la communauté scientifique. On peut citer l'un de ces premiers projets de grande envergure, le projet TCGA (The Cancer Genome Atlas), qui rassemble, pour 33 types de cancers différents, des données omiques multiples, comportant des données du séquençage de l'exome (partie codante du génome) et du transcriptome, des données de méthylation ainsi que des données cliniques pour chaque patient. L'analyse de données omiques introduit cependant plusieurs difficultés principalement dues à leurs grandes dimensions et à leur complexité biologique. Des méthodes computationnelles adaptées sont nécessaires pour analyser et interpréter ces données. Utilisant le cancer du poumon comme support d'étude, les travaux présentés dans cette thèse soulignent comment des approches intégratives peuvent améliorer l'analyse et la compréhension des données omiques en s'appuyant sur des méthodes telles que les méthodes d'apprentissage automatique. Le cancer du poumon est l'un des cancers les plus répandus dans le monde, avec environ 2 millions de nouveaux cas en

2018. Plusieurs types de cancers du poumon existent. Les plus courants sont généralement divisés en deux groupes : les cancers du poumon à petites cellules (CPPC) et les cancers du poumon non à petites cellules (CPNPC), qui représentent respectivement environ 20 et 75% des cas de cancers du poumon. Il existe également des formes plus rares de cancer du poumon. Certains de ces cancers ont été regroupés dans une catégorie, appelée tumeurs neuroendocrines (TNE) du poumon, par la classification 2015 de l'Organisation Mondiale de la Santé (OMS).

Les deux premiers chapitres de la thèse portent sur cette classe de cancer du poumon. Elle comprend les carcinoïdes pulmonaires, subdivisés en carcinoïdes typiques et atypiques, les carcinomes neuroendocriniens à grandes cellules (CNEGC) ainsi que les CPPC mentionnés précédemment. Les quatre types diffèrent sur plusieurs points. Les CNEGC et CPPC sont des carcinomes de haut grade, qui ont un mauvais pronostic et nécessitent des traitements agressifs ; tandis que les carcinoïdes typiques et atypiques sont, respectivement, des tumeurs de bas et moyen grade, qui présentent un meilleur pronostic (taux de survie globale à 5 ans de 82 à 100% et de 50% respectivement) et peuvent faire l'objet d'une résection chirurgicale. Etablir un diagnostic précis et correct de ces tumeurs est donc essentiel. Cependant, les critères actuels sont imparfaits, des variations de diagnostic entre pathologistes sont courantes et les marqueurs de diagnostics insuffisants.

Dans le premier chapitre de la thèse, nous avons effectué une caractérisation moléculaire des TNE du poumon, et plus particulièrement des carcinoïdes atypiques sous-étudiés. Des données multi-omiques, d'expression et de méthylation, ont été intégrées à l'aide de méthodes supervisées et non supervisées afin de mieux comprendre les différences et les relations entre les types de tumeurs et d'améliorer le diagnostic et la gestion cliniques de ces tumeurs. Un modèle basé sur la méthode du « random forest » (forêts d'arbres décisionnels), a été entrainé à prédire les sous-types histologiques des échantillons à partir des données transcriptomiques et du méthylome. Cette analyse supervisée a montré que la classification moléculaire ne correspondait pas exactement à la classification histologique et que les données moléculaires pourraient être utiles pour le diagnostic de ces cancers. Deux groupes de carcinoïdes atypiques, dont le prognostique diffère, ont été identifiés. L'un des groupes présentait une survie globale à 10 ans de 88%, similaire à celle des carcinoïdes typiques, et l'autre groupe une survie globale à 10 ans de 27%, similaire à celle du groupe des CNEGC. D'autre part, une analyse non supervisée a révélé l'existence de sous-groupes, caractérisés dans cette étude sur le plan moléculaire dans le but d'identifier des marqueurs avec d'éventuelles implications cliniques. Enfin, nous avons identifié le groupe des supra-carcinoïdes qui comprend des TNE avec une morphologie de type carcinoïde mais dont les caractéristiques moléculaires et cliniques correspondent aux CNEGC. Cette observation soutient l'hypothèse, précédemment proposée, d'un lien moléculaire entre les néoplasmes neuroendocriniens pulmonaires de bas et de haut grade.

Dans une seconde partie, nous avons généré une carte moléculaire des TNE du poumon en intégrant six jeux de données transcriptomiques et avons fourni de multiples ressources pour reproduire et étendre la carte moléculaire dans le futur. Afin de favoriser la réutilisation des données générées lors de l'étude présentée dans le chapitre précédent, les pipelines de

prétraitement à suivre et les points de contrôles qualité à valider ont été décrit pour permettre l'intégration avec d'autres données, précédemment publiées ou futures. En utilisant ces pipelines, nous avons harmonisé les données transcriptomiques de cette première étude avec cinq autres jeux de données. En appliquant la méthode de réduction de dimensions UMAP (Uniform Manifold Approximation and Projection) aux données homogénéisées, une carte moléculaire des TNE du poumon résumant l'expression de plus de 50,000 gènes en deux dimensions a été construite. Afin d'évaluer la qualité de cette carte, nous avons dans un premier temps vérifié qu'elle corroborait les groupes moléculaires identifiés et les hypothèses biologiques formulées dans les précédentes études. Dans un second temps, la préservation du voisinage des échantillons et des autocorrélations spatiales entre le jeu de données initial et la projection en deux dimensions a été estimée. Nous avons montré que l'intégration des jeux de données et la génération de la carte moléculaire permettaient de réidentifier les groupes moléculaires précédemment observés. De surcroit, les groupes identifiés par deux précédentes études indépendantes se sont avérés être les mêmes entités sur la carte moléculaire (superposition des échantillons des deux études sur la carte). Enfin, nous avons mis en avant diverses ressources afin de favoriser l'exploration de cet ensemble de données, la génération de nouvelles hypothèses, ainsi que l'intégration de données futures. En effet, les pipelines utilisés pour le prétraitement des données sont basés sur des langages et outils computationnels, tels que Nextflow, Docker et Singularity, favorisant la reproducibilité et la portabilité des analyses. Aussi, les données homogénéisées et le code, nécessaires à la reproduction de la carte moléculaire des TNE du poumon et des analyses effectuées qui en découlent, ont été mis à disposition sur GitHub et dans un journal computationnel interactif sur Nextjournal. Enfin, la carte moléculaire obtenue est en ligne sur TumorMap, un outil permettant une exploration interactive et statistique de cartes moléculaires comme celle proposée dans ce chapitre.

Le troisième chapitre de la thèse porte sur l'intégration de données constitutionnelles et somatiques au sein des cancers du poumon non à petites cellules (CPNPC). Ce groupe est subdivisé en deux sous-groupes principaux : les adénocarcinomes pulmonaires (lung adenocarcinoma, LUAD) et les carcinomes épidermoïdes du poumon (lung squamous cell carcinoma, LUSC). Jusqu'à présent, la plupart des études visant à comprendre l'étiologie et le développement de ces cancers ont concentré leurs efforts sur l'analyse des variations constitutionnelles ou sur les analyses de profiles somatiques de manière indépendante. Sur le plan constitutionnel, des études d'association pangénomiques (genome-wide association studies, GWAS) ont identifié plusieurs variants associés au cancer du poumon. Parmi eux, les variants liés à la consommation de tabac ont les effets les plus importants. Sur le plan somatique, les LUAD et LUSC ont été décrits, entre autres, comme faisant partie des cancers les plus mutés et présentant, pour les cancers associés au tabac, une signature mutationnelle associée aux dommages causés par les composants mutagènes de la cigarette et caractérisée par un excès de changements nucléotidiques de type C (base cytosine de l'ADN) vers A (base adénine de l'ADN) (généralement appelée Signature 4). Cependant, l'association directe entre les variants constitutionnels associés au cancer du poumon et la charge mutationnelle somatique dans les tumeurs du poumon n'a, à notre connaissance, pas été testée.

Pourtant, étudier les interactions entre les événements constitutionnels et somatiques pourrait d'une part, faciliter l'identification des gènes responsables de la susceptibilité au cancer du poumon et d'autre part, mettre en lumière les mécanismes de développement de ces cancers. En utilisant les résultats statistiques de GWAS rendus publiques ainsi que les données de génotypage et de séquençage de la base de données TCGA, nous avons étudié cette interaction à travers différentes approches. Tout d'abord, nous avons établi des scores de risque génétique (polygenic risk score, PRS) afin de combiner les effets de plusieurs variants constitutionnels en une mesure de risque de développer un cancer du poumon. Une association entre ces PRS avec le nombre total de mutations somatiques ainsi qu'avec le nombre de mutations attribuables à la signature mutationnelle 4, liée à l'exposition au tabac, a été observée. Cependant, cette association était principalement due aux variants liés au tabagisme et s'est avérée plus forte chez les LUAD que chez les LUSC. Afin de tester le lien causal entre la cigarette et la charge mutationnelle, nous avons utilisé la méthode de randomisation mendélienne qui permet, sous certaines hypothèses, de tester un effet causal entre une exposition et un phénotype en utilisant les variations génétiques, également appelées instruments génétiques, comme substitut de l'exposition. Bien que le lien de causalité entre l'exposition à la cigarette et la charge mutationnelle ait été confirmé par plusieurs tests de randomisation mendélienne, un effet pleiotropique a également été détecté par le test de Egger, ce qui suppose que d'autres mécanismes sont impliqués. Des analyses complémentaires évaluant l'influence de chaque variant étudié ont identifié un locus du chromosome 15q25, situé dans la région du gène CHRNA5 (sous-unité de récepteurs nicotiniques), comme étant responsable de l'effet causal détecté. Des analyses supplémentaires limitées à ce locus seraient donc nécessaires. En effet, la pléiotropie mise en évidence pourrait être liée à ce locus dont l'influence sur plusieurs phénotypes a déjà été caractérisée.

En conclusion, les travaux présentés dans ce manuscrit ont intégré différents jeux de données omiques afin de : i) explorer la diversité moléculaire des tumeurs neuroendocrines du poumon, ii) tirer parti de jeux de données transcriptomiques indépendants de TNE du poumon afin d'augmenter le nombre d'échantillons étudiés ainsi que contraster les profils moléculaires des tumeurs à travers la génération d'une carte moléculaire, et iii) explorer l'interaction entre la susceptibilité au cancer du poumon et la charge mutationnelle de ces tumeurs. Dans une dernière partie, nous proposons d'éventuelles améliorations et extensions des analyses décrites dans les différents chapitres et discutons les résultats présentés dans cette thèse dans le contexte de projets génomiques actuels et futures.

Acknowledgements

I would like to thank all the persons who contributed to this thesis:

The president Pr. Dumontet, the reviewers Pr. Relton and Pr. Thirlwell and the examiner Pr. Amos, for accepting to be part of my thesis jury and for the time they accorded to the evaluation of my work. I especially thank the reviewers for their insightful comments and suggestions on this dissertation. I also thank Pr. Amos and Dr. Scalbert for accepting to be part of my thesis committee and for their advices the last three years.

My supervisors, Dr. Matthieu Foll and Dr. James McKay. Thank you for welcoming me in the team and giving me the opportunity to work in the GCS group and at IARC. Matthieu, thank you for trusting me since the beginning, by accepting to supervise me as an intern and then as a PhD student. You transmitted your commitment for good science, I enjoyed all our discussions and learnt incessantly from you. Your patience and your continuous support allowed me to grow scientifically but also personally. James, thank you as well for having so much faith in me and for your patience. You pushed me and encouraged me at any time as well as transmitted your love of research by always reminding me to have fun. I have been very lucky to work with and learn from both you the last four years, it was a fantastic experience of which I will always keep great memories.

The Rare Cancer Genomics team, especially Dr. Foll and Dr. Fernandez-Cuesta for giving me the chance to work on their projects. Lynnette, you have been inspiring during my PhD, I truly appreciated you dedication, kindness and honesty. I also thank, in particular, Dr. Alcala, Ms. Mangiante, Dr. Leblay, Ms. Mathian and Dr. Voegele for being such amazing coworkers. I thank all of you for your contributions to Chapters 2 and 3, it has been a real chance to work with you all, both scientifically speaking and at the personnel level. Nicolas, you have always been ready to help, and discussing with you, whether it was on scientific topics or on my future, has always been very enriching. Lise, I particularly thank you for your listening, your humour and simply for your friendship. I wish you and Emilie all the best and a lot of success for the rest of your PhDs and careers.

Dr. Atkins for his contribution to Chapter 4 and for his precious support during this last year of PhD, thank you for encouraging me, pushing me and always bringing your positive thinking and energy. I thank you as well for proofreading this dissertation.

The GEN section, in which I met great scientists and colleagues from diverse backgrounds and from which I learnt a lot. I especially thank the members of the GCS group for their warm welcome. You provided me a wonderful working environment by always accepting to answer my questions, to help and to discuss. I have met in the group and

at IARC in general amazing colleagues and friends. I particularly thank Amélie, Thomas, Bertrand, Valérie, Matthieu, Lise, Nicolas, Karine, Noémie, Tiffany and Geoffroy, for all the moments spent at lunch, taking coffees and for the moments spent outside of work.

The Synergie Lyon Cancer team, especially Mr. Anthony Ferrari, Dr. Anne-Sophie Sertier and Dr. Alain Viari, for their supervision during my first master internship. Thanks to you, I discovered cancer genomics research in a fantastic team, that motivated me to pursue in the field. I would not have achieved this PhD thesis without you at the beginning.

My long-standing friends, Pauline, Nicolas, Romain, Diana, Guillaume, Nhi, Fanny and Olivia for their support during all these years at INSA and during my PhD. Having most of you in Lyon was a real chance and I hope this friendship will continue for a long time.

My family in Lyon area my aunt and oncle Isabelle and Nano, Coraline and Julien. Thank you for welcoming us all those week-ends, the afternoons of games and the wonderful Christmas in the snow. My family in Reunion Island, especially my parents and my sister, Emeline, for supporting me and encouraging me in every choices I made. Your support and love, even from 10 000 Km away, has been precious and has always been a huge source of motivation for me.

Yoan, for your listening, patience, unconditional support, encouragements and love since we met. Doing this PhD has not always been easy and your emotional support has been crucial in its success. I have been extremely lucky to have you by my side and look forward to our new adventures.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The biology of cancer

Cancer was the second cause of death worldwide, with almost 10 million deaths, in 2018 [1] and could in a near future become the leading cause [2]. The disease can affect different parts of the body, although some tissues are more frequently altered than others. Lung cancer, on which the work described in this manuscript will focus, is one of the most common cancers and the deadliest according to the 2018 GLOBOCAN database (a project of the International Agency for Research on Cancer (IARC) providing worldwide cancer statistics) [1]. Cancer is a complex disease that is highly controlled by the genome [3, 4]. It originates from normal cells whose genetic information has been altered. Those alterations can result from endogenous processes as well as from exogenous processes like environmental exposures and lifestyle [5, 6]. As a result of these alterations, tumor cells have acquired specific capabilities that allow them to grow in an uncontrolled way as opposed to normal cells. These capabilities are referred to as the hallmarks of cancer and are listed in Figure 1.1 [7].

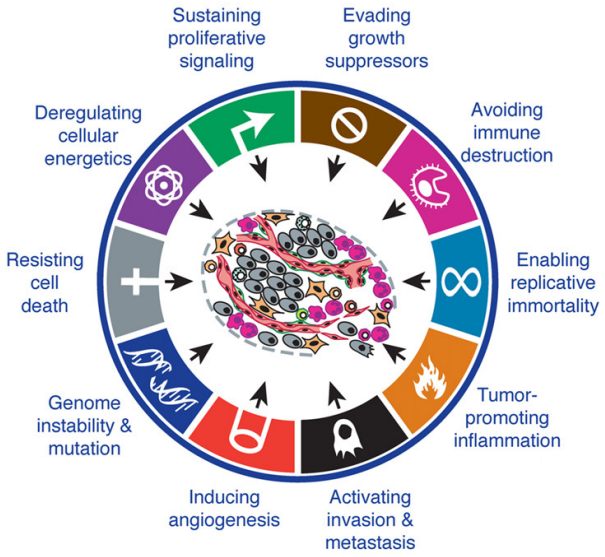


FIGURE 1.1: **The hallmarks of cancer**. From Hanahan *et al.* [7]

The first part of the introduction describes how genomic changes can influence cancer development and how the technological advances in the genomics area have enabled to shed lights on the mechanisms involved.

1.1.1 The central dogma of molecular biology

At the beginning of the 19th century, Avery and colleagues isolated and identified the Deoxyribonucleic acid (DNA) as the molecule constituting our chromosomes, defined previously as carriers of our hereditary material by Avery *et al.* [8, 9]. In 1953, Watson and Crick proposed a new structure for the DNA molecule, the double helix structure [10] (See Figure 1.2A). In 1968, Nirenberg *et al.* received the Nobel prize for interpreting the genetic code leading to protein synthesis. The process is described by what is called the central dogma of molecular biology (Figure 1.2B-C).

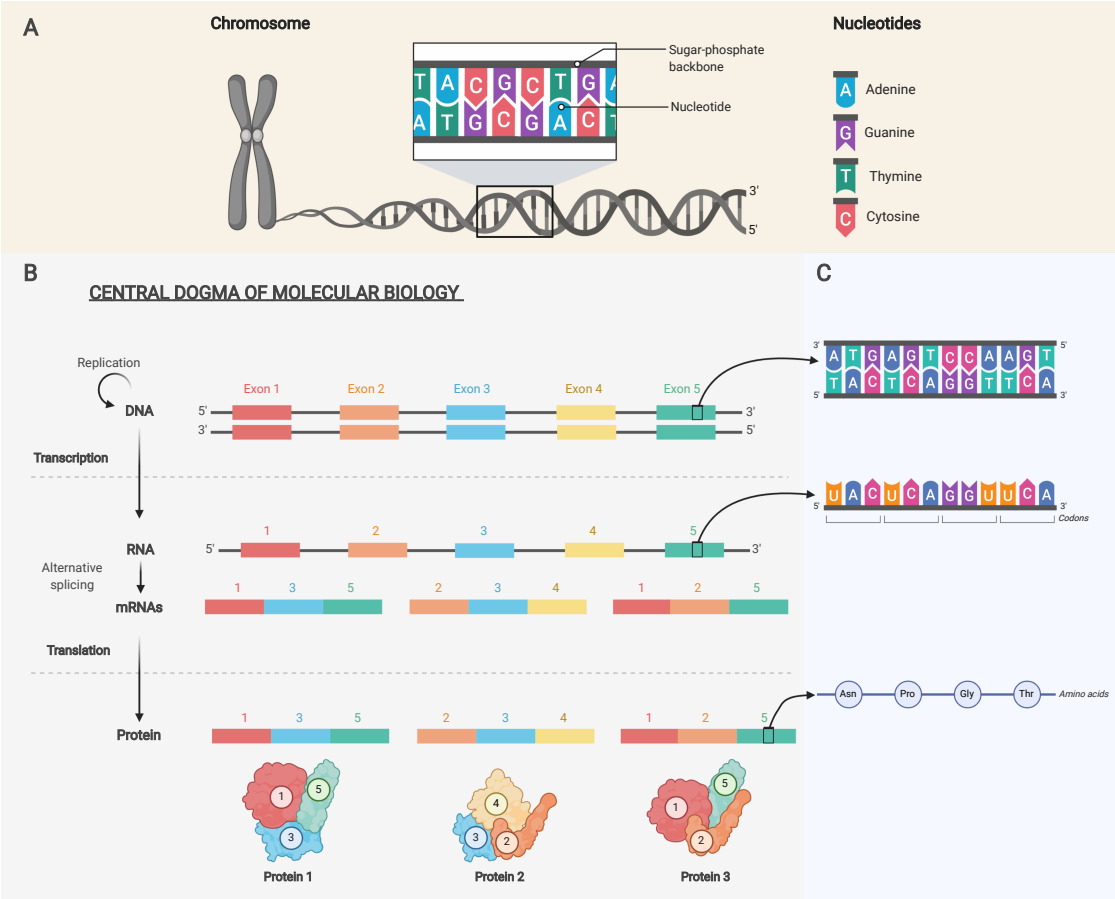


FIGURE 1.2: **The DNA molecule and the central dogma of molecular biology**. A) The structure of DNA: the double helix molecule is composed of two complementary strands of nucleotides. B) Representation of the steps described by the central dogma of molecular biology. C) Illustration of the molecules resulting from the central dogma transfers at a higher resolution. Created with BioRender.com

Three main transfers are described by the central dogma: replication, transcription and translation (See Figure 1.2B). During replication, the DNA molecule is duplicated to provide the needed information to progeny cells. Through the two other steps, the information contained in DNA is used to generate proteins. Firstly, the process of transcription consists in reading the DNA sequence to synthesize a single-stranded molecule of the same length, the Ribonucleic acid (RNA). During translation, the transcribed molecule is then read using a reading frame of three nucleotides that form what is called a codon encoding for one amino acid, the unit of a protein (See Figure 1.2C). Note that the genetic code is redundant; multiple codons can encode an amino acid. The conversion of the information encoded in our genes to functional gene products like proteins is referred to as gene expression.

Since the statement of the central dogma, other mechanisms have been identified as determinant for the expression of a protein. Firstly, the RNA molecule resulting

from the transcription process, containing regions coding for the final amino acids sequence (exons) and non-coding regions (introns), is actually a precursor messenger RNA (pre-mRNAs). The step transforming precursor RNA to mature messenger RNAs (mRNAs) is called splicing and consists in truncating intronic regions and joining different exons together (See Figure 1.2B). One pre-mRNA can lead to multiple mRNAs that are then transported outside of the nucleus to be translated into different isoforms. While around 20,000 genes are described, much more proteins are generated as a result of this process called alternative splicing.

Although all our cells share the same genetic information and follow the same dogma, it is known that cells in distinct tissues differentiate and do not express the same proteins, at the same time. Such differences can be explained by the fact that several regulatory processes control gene expression levels. Firstly, genes transcription is dependent on transcription factors that represent around 7% of the genes [11]. They specifically bind to control regions of genes, provide or prevent access to the DNA and can control multiple genes [11]. The fact that genes, for example the transcription factors, can influence multiple genes and thus sometimes multiple unrelated phenotypes is referred to as pleiotropy. After transcription, mRNAs can also be regulated through other RNA molecules, like the micro RNAs (miRNAs), that can degrade mRNAs. Besides, differences in gene expression can be controlled via non-genetic mechanisms like epigenetic processes, including histone modifications and DNA methylation. Histones are proteins around which the DNA is wrapped and hence control DNA accessibility (Figure 1.3). For example, histone phosphorylation leads to the condensation of the chromatin and inhibits gene expression [11]. DNA methylation consists in the addition of a methyl group to cytosine nucleotides located in cytosine–phosphate–guanine (CpG) dinucleotides sites (cytosine followed by a guanine nucleotide). Such positions are not homogeneously distributed across the genome and are more frequently observed in what is called CpG islands, themselves mainly observed in regulatory regions of genes, the promoters. It has been observed that the methylation of CpG sites in promoters can repress gene expression while methylation of positions in the gene body positively correlates with gene expression [12].

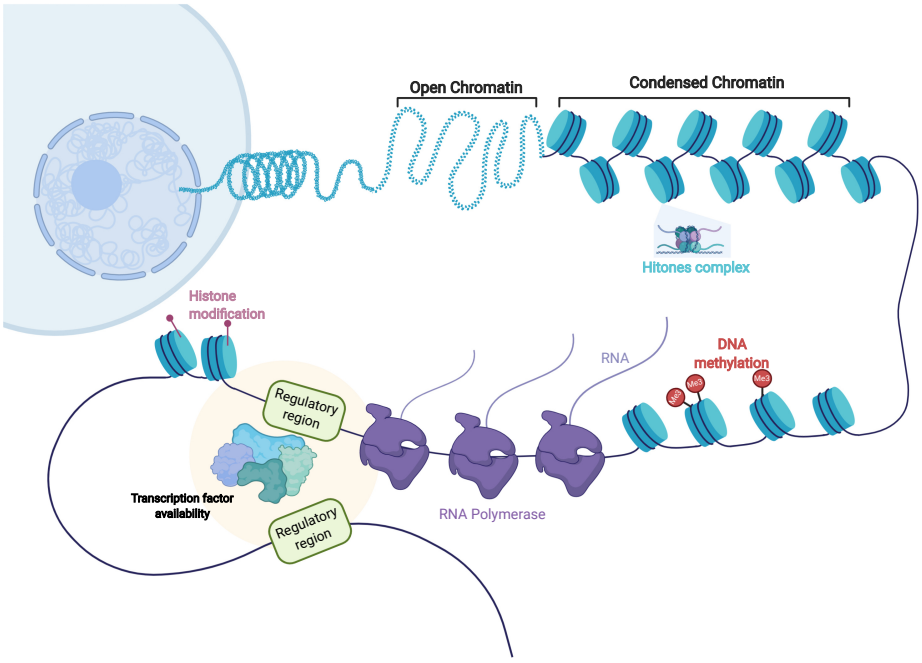


FIGURE 1.3: **Regulation of transcription**. The figure represents different configurations of DNA packaging. The DNA molecule is wrapped around histones proteins that themselves are gathered in complexes called nucleosomes. This packaging forms the chromatin structure. This structure can be more or less compact (open versus condensed chromatin), which is influencing gene expression. When the chromatin is open, transcription factors can access the DNA molecule, and RNA polymerases can initiate the transcription. Note that the structure of the chromatin can be influenced by histones modifications and DNA methylation events. Created with BioRender.com

Finally, post-translational events like enzymatic modifications of proteins or protein cleavage can occur and affect proteins functions, hence adding an additional layer of complexity.

As such, the numerous steps of transferring the DNA sequence information to proteins reflect the complexity behind protein expression. Any of these steps can be disrupted and result in altered molecules and proteins, leading to cancer development.

1.1.2 Cancer: a genomic disease

Our DNA continuously undergoes diverse alterations and their accumulation over time can cause cancer. Researchers started to investigate the role of genomes in cancer at the end of the 19th century. In 1890, David von Hansemann, by observing cancer cell division under a microscope, identified for the first time abnormal chromosomes. This observation, among others, led Theodor Boveri 20 years later

to suggest that cancer was a consequence of alterations in our inherited DNA [3]. His hypothesis was supported in the mid 20th century by the identification of a recurrent alteration resulting in a peculiar chromosome 22 (the Philadelphia chromosome), in chronic myelogenous leukemia (CML). While those alterations have been observed at the chromosomal level, genomes can be impacted by a multitude of alterations detectable at a higher resolution, the modification of one nucleotide in the DNA sequence being the highest resolution.

At any position of the genome, the nucleotides might vary from an individual to another as well as between cells of an individual; those variations are called Single Nucleotide Variations (SNVs). Also, larger events like nucleotides insertions or deletions (indels) of up to 1,000 bases and structural variations (chromosomal rearrangements or large indels) can alter the DNA sequence. All of these genomic changes are called mutations.

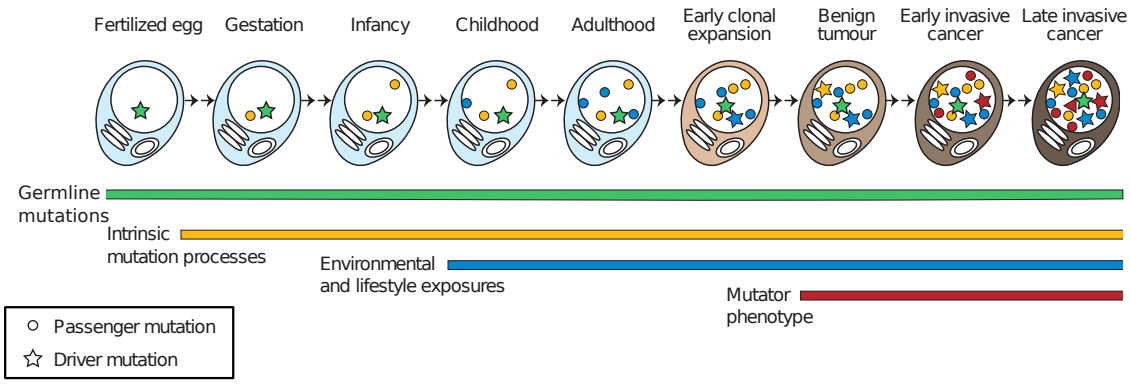


FIGURE 1.4: **The timing of somatic mutations acquisition**. Mutations can be inherited at birth (germline mutations, in green) or acquired during life course (somatic mutations, in yellow, blue and red). They can have little to no impact (passenger mutations represented by circles) or confer an advantage to the cell (driver mutations represented by stars). Adapted from Stratton *et al.* [3]

Mutations can occur at different moments in life (See Figure 1.4). Some mutations are inherited at birth since they are present in the germ line cells (sperm and egg) transmitted by parents to the offspring. They are called germline mutations and are found in all the cells of an individual, normal cells as well as tumor cells. Such mutations are observed at different frequencies in different populations and are called Single Nucleotide Polymorphism (SNP)s. Another category of mutations can also be found in all cells of the body even if they were not transmitted by our parents, if they occur early in life during the development, at gestation. They are

called *de novo* mutations. Finally, the rest of the mutations found in humans are acquired later in life as a result of errors in the DNA maintenance or exogenous damages (See next section). Those mutations occur in cells outside the germ line and are called the somatic mutations.

Also, whether they are germline or somatic, mutations can have different impacts. Most mutations have, due to the redundancy of the genetic code or due to their location in the genome (*i.e* in non-coding regions), little to no impact on the genes encoded around them, they are the passenger mutations [13]. Others though alter the gene product or its expression levels and confer a selective advantage to the cell, *e.g.* a faster proliferation or a better survival in comparison to neighbour cells [3]. Those mutations are called driver mutations as they are thought to contribute to "driving the carcinogenic process" and are preserved by positive selection. In 2018, the Cancer Gene Census described more than 700 driver genes (genes carrying driver mutations). Among them, 90% were associated with somatic mutations and 20% contained germline mutations [14, 15]. Generally, two types of driver genes exist, oncogenes and Tumor Suppressor Genes (TSG). Oncogenes are genes whose functions are thought to promote cell growth, proliferation or inhibit apoptosis and usually result from a gain of function. A mutation in an oncogene can thus lead to a dysregulation of one of these processes, hence resulting in uncontrolled proliferation and cancer. The first mutation identified as causing cancer was discovered in 1982 by Reddy *et al.* and activates an oncogene named *HRAS* [16]. Besides mutations, other processes like over-expression of genes via amplification or chromosomal translocations can activate this category of genes. In contrast to oncogenes, TSGs are restraining cellular growth and proliferation and are often referred to as the "gatekeepers" genes. Mutations in TSGs tend to result in a loss of function; the latter genes are inactivated, and their negative regulation of cell proliferation is cancelled, which leads to abnormal growth. In 1971, Knudson proposed the two hit hypothesis, which stipulates that both alleles (versions of a gene inherited by our mother and father, identical alleles leading to the homozygous state while two different alleles to the heterozygous state) of a TSG must be inactivated or lost for the gene to lose its normal functions [17]. This hypothesis seemed to explain some familial cases such as retinoblastoma or Wilms' tumor [18]. Indeed when the first hit is an inherited germline mutation, the cancer susceptibility of a person increases since only one alteration is needed to alter the TSG functions. The second alteration can result from different events: a mutation in the second allele, the loss or translocation of chromosome pieces or the loss of an entire chromosome. The two latter events causing what is called loss of heterozygosity (LOH) [5].

In the case of the two hit hypothesis, two mutations in the same gene are required for cancer initiation. However, it has been described that cancer usually results from a multi-step process, meaning that multiple mutations and more than one gene are usually involved. A certain number of alterations in key pathways are necessary, and it can take several years for cancer to develop [11]. However, the multi-step process can be accelerated. Firstly, as mentioned previously, the inheritance of germline mutations speeds up the cancer development as one driver mutation might be present from birth, increasing the probability that the remaining necessary events, which generally follow a stochastic process, will also occur. [11]. Also, even if multiple DNA repair mechanisms fix most of the alterations that a genome endures, the DNA repair pathways themselves can be disrupted, leading to an acceleration of the accumulation of alterations. Such an event increases the mutation rate of an individual and generates what is called a "mutator phenotype" [3, 19]. Finally, driver genes can also be altered by epigenetic changes that are more frequent. Such changes increase the chance of disrupting key biological pathways for cancer development.

1.1.3 Cancer: an environmental disease

Mutations can arise from endogenous processes, for example, errors happening during DNA replication. In that regard, the appearance of mutations across the genome seems random, and the advent of a driver mutation leading to cancer development seems associated with bad luck. This idea has been developed by Tomasetti *et al.* [20] in a controversial paper, published in 2015, suggesting that the majority of cancer mutations were due to "bad luck". In 2017, the same authors confirmed that mutations due to random errors represent a large proportion of mutations in multiple cancers while specifying that if luck and randomness do play a role in cancer development, other factors like exogenous processes also impact our DNA and contribute to cancer development [21]. Note that this study is, for some cancers, in contradiction with the work of a study estimating intrinsic risk of different cancers and being critical of the initial controversial paper [22].

Cancer incidence varies depending on the countries considered. Lung cancer incidence, for example, is much higher in Asia, Europe and North America than in Africa [23]. Those differences can be explained by the fact that cancer has a heritable component that differs in different parts of the world and by the fact that environmental exposures are different across countries. It has been shown, though, in studies exploring cancer rates in migrants populations, that the differences observed among populations could not be explained only by the genetic component

[24]. In the second half of the 20th century, epidemiological studies have indicated that several environmental exposures were associated with cancer incidence, showing that many cancers could be prevented. One of the most striking findings was that of Doll *et al.* showing that smokers had a twenty-fold higher risk of developing lung cancer than non-smokers [25]. At the same period, chemical agents have been identified as being able to induce cancer, *i.e.* being carcinogenic [26]. Some of these agents were also defined as mutagenic agents, *i.e.* agents inducing DNA damages.

Some carcinogens can impact cancer evolution without causing DNA alterations; they are non-mutagenic agents and are considered as tumor promoters. One example of tumor promoter is alcohol which is a cytotoxic substance. Its consumption leads indeed to the death of epithelial cells in the mouth and throat, which triggers the division of the stem cells to regenerate the epithelium. If tobacco consumption precedes this event, tobacco-induced mutations might be present in the dividing cells, and clonal expansion of these mutations may lead to cancer [11]. In that case, smoking acts as a tumor initiator and alcohol as a promoter by stimulating cell proliferation. Such interaction between alcohol and smoking is observed in head and neck cancers. Note, however, that alcohol can also have a mutagenic effect due to metabolites, like acetaldehyde, generated during ethanol oxidation in the liver [27]. Other examples of tumor promoters are steroid hormones acting as mitogenic agents or chronic inflammation (*e.g.* due to viruses).

We have seen that mutations in our genome can result from endogenous processes like replication errors or DNA repair defects and from exposition to carcinogens. Observing these mutations across the whole genome have revealed patterns. Indeed, each of these processes can generate what is called mutational signatures, *i.e.* specific combinations of mutations [28]. The first studies of mutational signatures focused on single base nucleotide substitutions (six possible substitutions: C>A, C>T, C>G, T>A, T>C, T>G) and their tri-nucleotide contexts (the 5′ and 3′ nucleotides flanking the substitution) leading to 96 possible classes of mutations. The classification of all mutations found in cancer genomes in those 96 groups and the use of mathematical methods (See section 1.4) to decompose the mutational processes enable the identification of a limited but diverse set of signatures. In the case of lung cancers, comparing the DNA of smokers with that of non-smokers revealed an increase of mutations in smokers mainly due to an elevation of C to A (C>A) mutations, probably caused by the tendency of tobacco carcinogens to induce this particular change [29]. In melanoma samples, an increase of C>T substitutions has been identified as a result of Ultraviolet (UV) light exposition [30]. In 2015, COSMIC provided a curated set of 30 mutational signatures based on previously published

studies on different cancer types [31]. Recently the methods to disentangle mutational signatures in human genomes have been extended. In 2020, Alexandrov *et al.* have considered higher context to classify single base substitutions by considering two flanking bases around the positions of the mutations and analyzed as well other types of mutations like double base substitutions and indels. This work led to an expansion of the repertoire of mutational signatures with more than 60 signatures in total [32].

Although some signatures are resulting from endogenous processes, like defects in DNA repair or unknown processes, multiple signatures have been associated with preventable exposures. Considering the important impact of environmental exposures, Wild *et al.* suggested in 2005 the concept of the *exposome* which corresponds to all the exposures encountered by an individual during his lifetime (*e.g.* life-style, exposures to chemicals). He expressed the need to improve the measurement of such exposures at the same scale of the genomic events measurements [33]. Indeed on the genome side, remarkable technological advances were made in the past decades allowing researchers to explore the human genome at high resolution. The evolution of these technologies is described in the next section.

1.2 The era of genomics

1.2.1 From arrays to next generation sequencing

The identification of the genomics variations leading to cancer has been enabled by multiple technical and technological advances that occurred after the discovery of the DNA structure. Since that discovery, researchers have attempted to decipher the hidden information contained in the double helix molecule. One fundamental advancement in genomics has been the development of the first generation sequencing by Frederic Sanger in the 1970s. After automatization, this technique led indeed to the sequencing of the first human genome in the context of the Human Genome Project (HGP) that started in the 1980s, took 13 years and cost around 3 billion dollars to lead, in 2003, to the sequencing of the 3 billion nucleotides that our DNA constitutes. At that time, the largest genome sequenced was the 20,000 times smaller genome of the Epstein-Barr virus [34]. While many researchers thought it was impossible, the project completed and delivered the first version of the human genome reference which, after being revised and improved, is now used on a day-to-day basis in genomics. However, the first generation sequencing technology was

too long and costly to be applied in larger research projects aiming in that period to catalogue the genetic variations involved in human diseases.

The array technology

At the same period, the microarrays technologies were far less expensive. This technique consists in disposing, on an array, DNA sequences, called probes, designed to bind (by hybridization) to target sequences in a sample. The target sequences are labelled to measure the hybridization and quantify the target molecules.

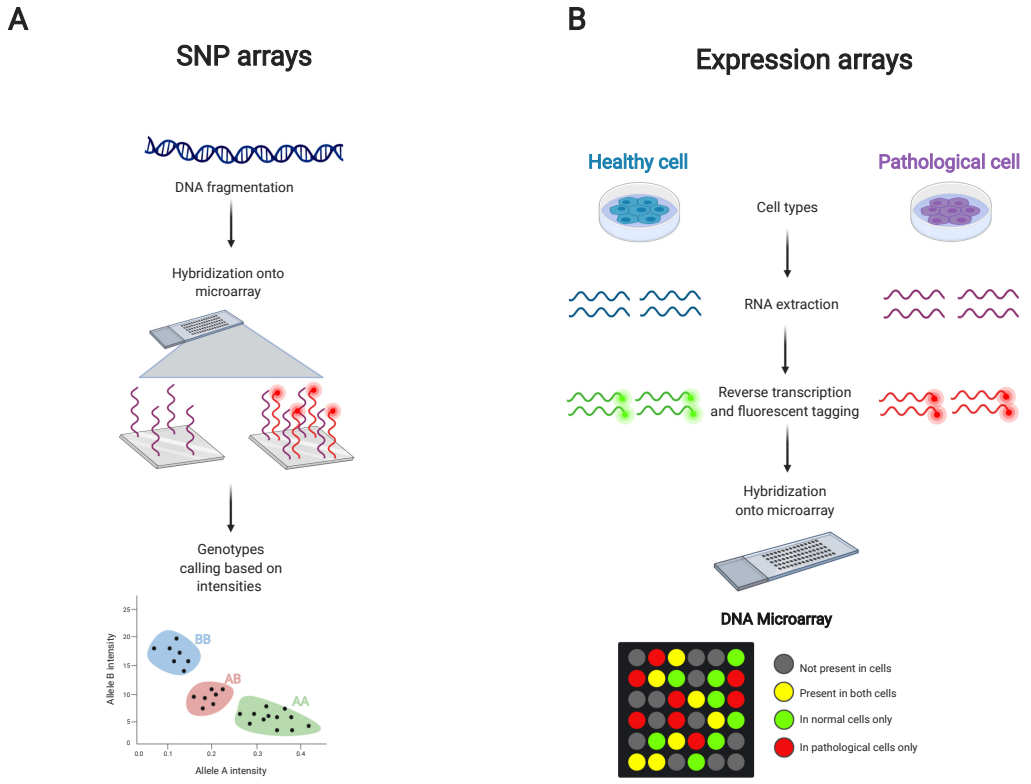


FIGURE 1.5: **Microarrays**. A) SNP arrays: fragmented DNA sequences bind to designed probes on the microarray, which generates an intensity signal that varies depending on the allele carried by the DNA sequences. B) Expression arrays: tagged complementary DNA, reverse-transcribed from mRNAs molecules, bind to gene-specific probes, which generates a fluorescence signal used to compare expression levels in different cell conditions. Created with BioRender.com

In order to study genomic variations across the genome, specific microarrays were developed, the genotyping or SNPs arrays. Those arrays contain unique probe sequences, targeting specific positions of the genome, which hybridize to single-stranded DNA that has been fragmented. This generates intensities signals varying

depending on the allele carried by the DNA sequence binding to each probe. This intensity, indicating the presence or absence of each allele, is then converted into genotypes [35] (See Figure 1.5A). The SNP arrays developed for commercial purposes have evolved, interrogating from 10,000 to millions of sites simultaneously in a given individual [36]. Key produces of these technologies were developed by Affymetrix and Illumina inc. Those arrays have been used so far for different purposes. They allowed the identification of copy number changes or, for arrays with high marker density regions, the detection of LOH events by identifying regions without heterozygous positions [37, 38]. They have also been used to identify germline variants that associate with a certain disease through Genome-Wide Association Studies (GWAS) [39]. As illustrated in Figure 1.6, GWAS interrogate millions of positions across the genome by testing their association with a specific trait, like smoking traits, individually and reveal positions significantly associated with that trait.

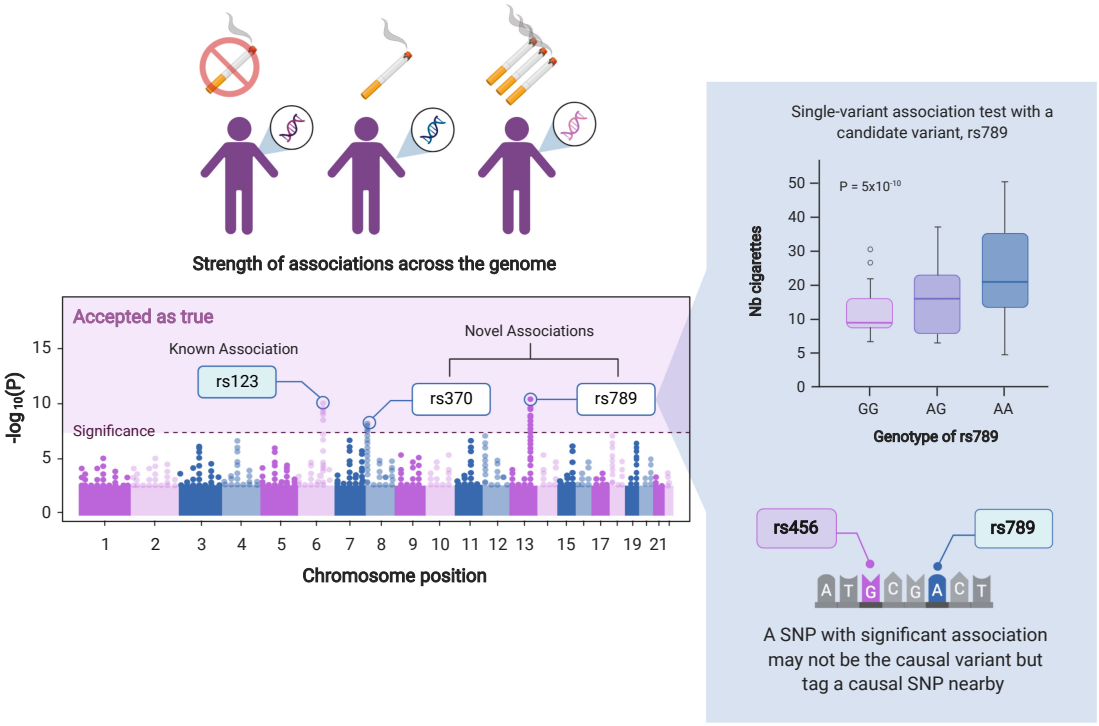


FIGURE 1.6: **Genome-wide association studies**. The figure illustrates a GWAS identifying SNPs associated with the number of cigarettes smoked per day. For each position, the association between the variant genotypes and the number of cigarettes per day is tested (rs789 example). The associations *p*-values are represented in a Manhattan plot (left panel). SNPs reaching the genome-wide significance threshold of 5.10^{-8} are considered as true associations. Those SNPs do however not always correspond to the causal variant but often tag a nearby SNP in linkage disequilibrium. Created with BioRender.com

Although SNPs arrays are limited to the positions assayed, much more positions can be studied based on the arrays. Indeed, SNPs are transmitted to the offspring linked to other close SNPs in blocks called haplotypes. This correlational relationship between SNPs is called linkage disequilibrium (LD). Knowing the SNPs composition of a haplotype enables to predict the genotype of SNPs that were not assayed by the array by using the information of the assayed positions in the haplotype. Hence, genotyping hundred thousands of SNPs allows actually to impute the genotype of millions of other variants thanks to LD. The definition of the haplotypes required though to study such genomic structure in different samples to build a map as reference. Those were the goals of the Haplotype Map project (HapMap) started in 2002 [40, 41].

Micro-arrays platforms have also been used to study the other molecular layers

like the transcriptome and the methylome. For the analysis of the expression profile, micro-arrays have enabled to measure and compare the expression levels of specific genes in cells under different conditions, *e.g.* diseased versus healthy cells or treated versus non-treated cells. Figure 1.5B describes the main steps of an expression array experiment. The extracted mRNAs molecules from both types of cells, after being reverse-transcribed to complementary DNA (cDNA) and labelled with fluorescent dye, hybridize to the genes specific probes fixed on the array. The array is then scanned using fluorescent imaging [42]. The fluorescence amount detected at each probe is proportional to the amount of mRNAs in cells. While these measures do not provide absolute quantification of gene expression levels, they enable to compare the expression levels in the different conditions.

Arrays have also been used to study the epigenome by allowing the detection and the analysis of methylation events. The most commonly used methylation arrays are the Illumina arrays [43]. Probes are designed to target specific loci of the human genome, CpG positions. The number of positions interrogated by such arrays can vary from 25,000 to 850,000 positions depending on the array (*e.g.* Illumina 25K, 450K and 850K arrays). Probes are designed and fixed to the array to bind to both methylated and unmethylated loci (Figure 1.7). This binding is enabled by a chemical process called bisulfite conversion, which converts unmethylated cytosines to uracil and leave methylated cytosine unchanged. At the hybridization step, a single-based extension is performed with labelled nucleotides, allowing to distinguish for each locus a methylated vs non-methylated signal (Figure 1.7). The ratio between the two signals at a locus provides a value, called β value, which indicates the level of methylation. This value ranges between 0 and 1, 0 corresponding to a non-methylated and 1 a methylated position.

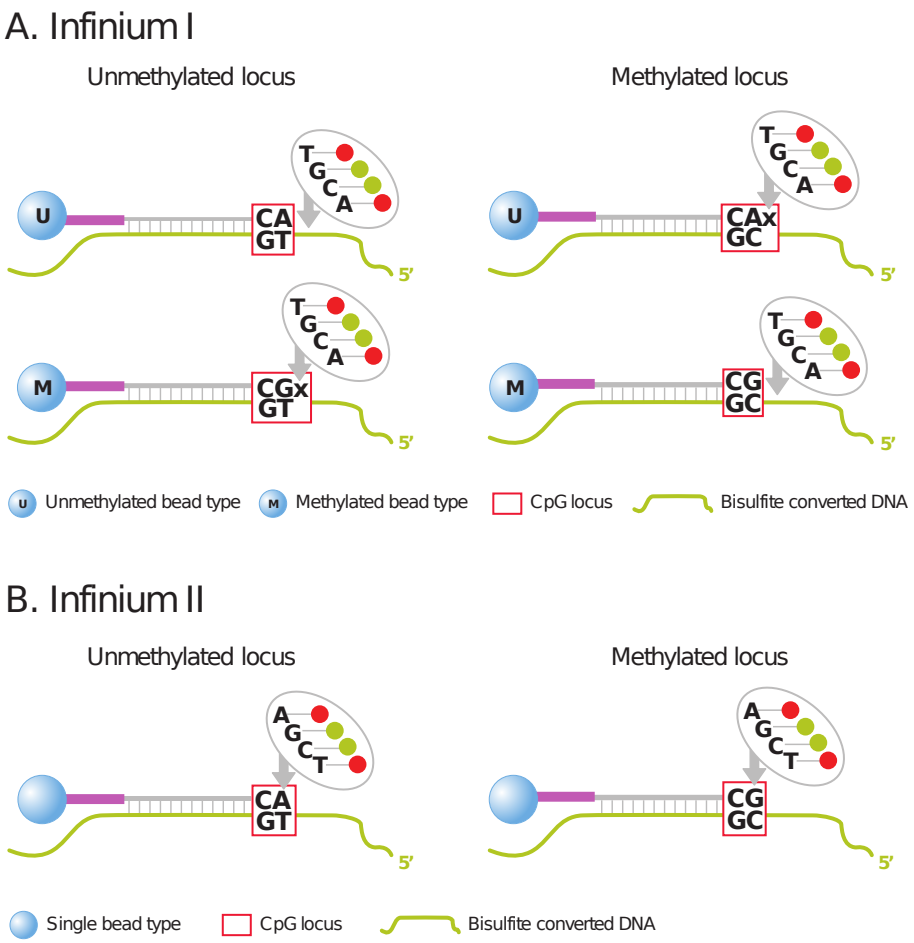


FIGURE 1.7: **The Illumina Infinium methylation assay** (From [43]). This figure represents the probes used for methylation profiling by Illumina. A) Infinium type I probes. Two site-specific probes are found on the array: probes allowing methylated sites with the preserved cytosine to bind (methylated bead M) and probes designed for the unmethylated site with the thymine nucleotide resulting from bisulfite conversion and whole-genome amplification (methylated bead U). B) Infinium type II probes. Only one probe per locus is required to bind to both methylated and unmethylated sites. In that case, single-base extension with labelled nucleotides is used.

Next-generation sequencing

While the SNP arrays enabled to access the genotype information of millions of positions, there was still a need to re-sequence human genomes more efficiently and access the complete DNA sequence to better identify genetic variations. Around 2005, the second generation of sequencing methods called Next Generation Sequencing (NGS) has been developed.

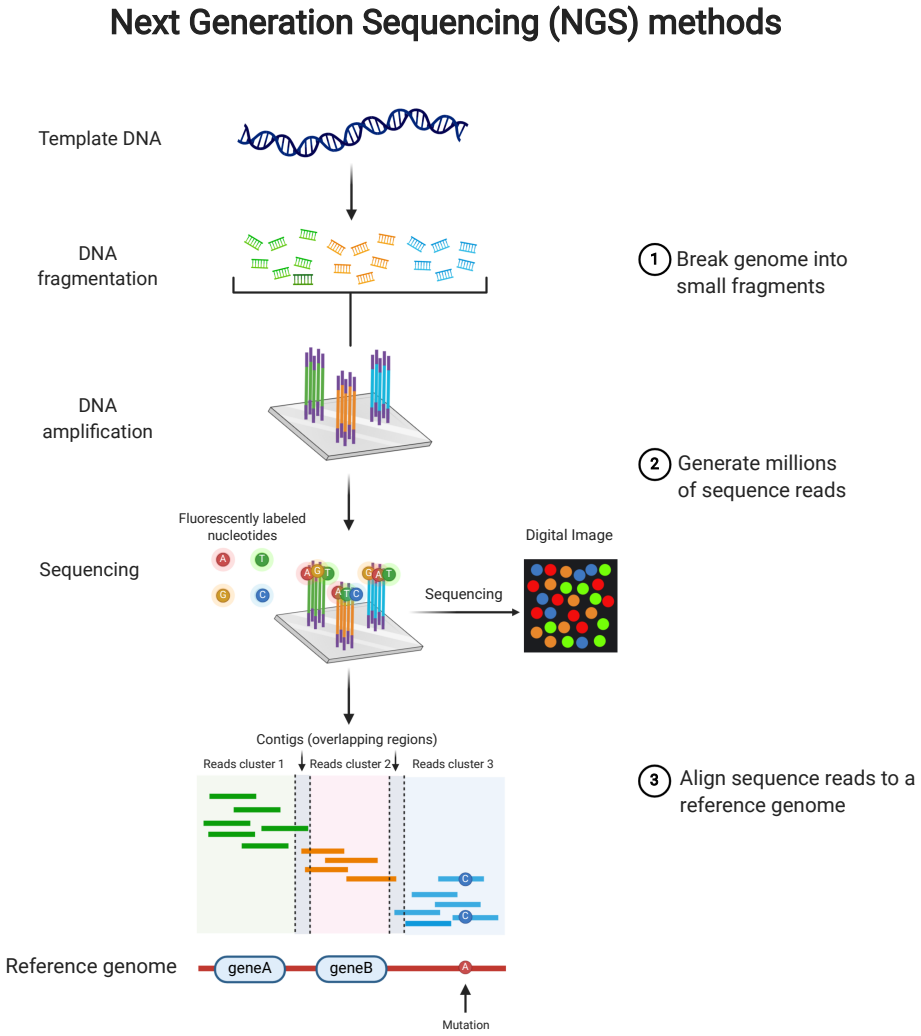


FIGURE 1.8: **Next Generation Sequencing methods**. The figure describes the NGS steps consisting in: i) fragmenting the nucleic acid molecule, ii) amplifying the fragments (using Polymerase Chain Reaction (PCR)), iii) sequencing the resulting copies using single-base extension that adds one after the other labelled nucleotides whose signals are detected using digital imaging. The sequencing reads are then aligned to a reference genome to assemble the reads in a single sequence or to detect mutations across the genome. In the case of RNA sequencing, the reads align to exonic regions of the genes and they are counted to quantify gene expression levels. Created with BioRender.com

The main change in these new methods in comparison to the first one was the parallelization of the sequencing, which allowed to produce millions of sequences, called reads, at the same time and hence to decrease drastically the time of sequencing as well as its cost [44] (Figure 1.8). NGS methods enabled the rapid re-sequencing of different parts and lengths of the genome. The entire genome sequence (except some highly problematic regions) can be accessed with Whole Genome

Sequencing (WGS). The restricted sequencing of coding regions (exonic regions) can be performed with Whole Exome Sequencing (WES). Finally, it is possible to sequence specific regions of the genome, usually genes, using targeted sequencing. Based on these techniques, bioinformatics methods have been developed to detect germline as well as somatic variants. They consist in mapping (or aligning) the sequenced reads to a reference genome, and positions that vary from the reference are identified as variations (Figure 1.8). A mismatch between a sequenced genome and the reference genome is expected around every 1,000 bases. To distinguish somatic from germline mutations, both tumor and normal cells DNA from the same individual have to be sequenced. The tumor DNA is compared to the normal DNA and variations found in the tumor cells only are classified as somatic mutations. Somatic mutations are expected every 1,000,000 bases approximately depending on the cancer type [28].

While the DNA sequencing techniques have been used to detect DNA mutations, they do not explore the expression or methylation layers. In 2008, the sequencing of the RNA molecule (RNA Sequencing (RNA-Seq)) had been performed to study expression profiles. In this technique, the mRNAs molecules are fragmented and converted to complementary DNA before sequencing, and the resulting reads are aligned to the reference genome [45]. After the alignment step, the reads can be assigned to genes and the abundance of reads mapped on a gene, quantified using the number of mapped reads, reflects the expression level of the gene (Figure 1.8). A high read count value indicating that a gene is active and transcribed in that sample. The comparison of the read counts distributions in samples from different conditions, *e.g.* samples with and without disease or diseased samples under different treatment, can be used to identify genes involved in or causing a specific condition. RNA-Seq can also be used to identify different transcripts of a gene as well as gene rearrangements like translocations.

Note that other recent techniques, while not described in the thesis, also exist to access different omics layers. A new sequencing technique has been developed for the analysis of the methylome, the bisulfite sequencing, which in contrast with the methylation arrays, can interrogate millions of CpGs positions across the whole genome as well as positions in targeted regions. Also, the study of chromatin accessibility and DNA-binding proteins is possible thanks to Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) and Chromatin immunoprecipitation experiments followed by sequencing (Chromatin immunoprecipitation Sequencing (ChiP-Seq)) respectively [46, 47]. Finally, while the sequencing methods presented so far process DNA coming from a bulk of cells, single-cell sequencing methods

have been developed to perform molecular characterization at the cell level. These methods allow the identification of distinct populations of cells in a tumor and hence the study of tumor heterogeneity and tumor microenvironment [48, 49].

The decreasing costs of genotyping and sequencing methods have enabled the establishment of genomics studies involving large cohorts [44]. Sequencing a human genome today costs less than 1,000 dollars using NGS methods while it would still cost millions if the Sanger method was chosen. Multiple research groups have coordinated their efforts to create large consortia for that purpose and in many cases have shared the resulting data to the scientific community. The next section provides an overview of some of these initiatives.

1.2.2 Large public databases

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a public database providing access to 10,000 patients whose tumors have undergone multi-omics characterization. The project was launched in 2005 by the National Institutes of Health (NIH) and aimed at characterizing the genomic alterations underlying several cancer types. For that purpose, multiple omics data were generated [50]. The tumor and normal samples from most of the TCGA participants have been sequenced using WES. Based on these data, multiple variant callers have been used to catalogue the germline and somatic mutations present in each sample. Genotyping has been performed to analyze copy number variations. The transcriptome of most samples has also been sequenced, using RNA and miRNAs sequencing. The methylation profiles of the tumors were explored with the use of 25K or 450K methylation arrays. Finally, protein expression profiling has been performed based on Reverse-Phase Protein Array (RPPA). In addition to the molecular data, clinical and environmental exposures data were collected when possible. The TCGA projects also delivered the histopathological images associated to each tumor. Based on these diverse omics and clinical datasets, "marker papers" describing the molecular landscape of each tumor type have been published. While the tissues explored at the beginning of the initiative were limited to lung, brain and ovaries, the TCGA data encompass today molecular data from 33 different cancer types. Those cancer-specific studies led to the identification of genomics alterations causing each cancer type, hence the discovery of new driver genes and potential cancer biomarkers, *i.e.* molecules found in the body as an indicator of a disease or specific condition. Also, cancer subtypes were characterized

on the molecular level and subtype-specific alterations were identified, which resulted in new clinical managements of tumors [51]. In parallel to the cancer-specific studies, the TCGA research network launched, in 2012, the Pan-Cancer Atlas initiative aiming at exploring the commonalities between cancer types, distinguishing tissue-specific determinants of cancer as well as increasing the statistical power for the identification of genomic alterations [51]. This initiative was completed in 2018 and the data have been released and associated to 27 papers, published in Cell, focusing on three main topics: i) cell-of-origin patterns and cancers subgrouping, ii) oncogenic processes, and iii) signaling pathways involved in cancer [52].

The International Cancer Genome Consortium (ICGC) initiatives

The TCGA studies focused their efforts on the characterization of the cancer exomes. However, exomes represent only 1% of the human genome and much more can be discovered by exploring the remaining 99% of the genome. In 2007, the International Cancer Genome Consortium (ICGC) project was launched to study more than 20,000 whole genomes from 50 cancer types having an impact in multiple regions of the world (the 25k initiative). The international consortium aimed at generating a catalogue of the somatic mutations in those cancer types, sharing the resulting datasets and complementing them with transcriptomic and epigenomic datasets [53, 54]. Based on the samples included in the TCGA and the ICGC projects, the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, an ICGC initiative also know as the Pan-Cancer project, has arisen [55]. The project relied on more than 2600 samples from 38 different tumor types and aimed at meta-analyzing whole-genome data across cancers along the same lines as the PanCancer Atlas project. The first results from these data have been released in 2020 in a series of publications in Nature [54]. While the TCGA initiative enabled the study of the coding regions of the samples, the PCAWG project, thanks to the use of whole genome sequences, was designed to explore broader mutational patterns in the coding and non-coding regions, from small to large events like structural variations. For example, chromoplexy and chromothripsis events, which are complex chromosomal rearrangements resulting from catastrophic genomic events, have been observed in more cancers than expected, 17.8% and 22.3% of the tumors, respectively [54]. Also, one major result from the PCAWG project has been the expansion of the mutational signatures mentioned in section 1.1 [32], as well as the discovery of 16 structural variants signatures [56].

UKbiobank

The previously described projects mainly targeted the somatic landscape of genomes. Other large projects have enabled the research community to explore the germline component of human disease. The largest public dataset, focusing on germline genetics, has been generated by the UKbiobank project, which started in 2010 in the UK. This project gathered data from a population-based cohort of around 500,000 participants between 40 and 69 [57] and had as main objective to improve our understanding of the interaction between genetics and multiple human diseases. For that purpose, all participants were genotyped. Besides, multiple other biological samples, like urine, blood and saliva as well as physical measures, *e.g.* brain Magnetic Resonance Imaging (MRI), heart and eye measurements, were collected. It is a prospective cohort; participants are followed up and are linked to electronic health records [58]. The genotyping data of the full cohort were released in 2017. Based on this dataset and the large panel of phenotypes, a multitude of GWAS studies related to human diseases have been performed and their resulting summary statistics were made available. In 2019, around 100 GWAS studies resulting from the UKbiobank data were available on the GWAS catalogue, which provides curated GWAS summary statistics results [59]. The follow-up of the patients has established that, in 2018, 79,000 of the participants were diagnosed with cancer [58], which means that cancer-related traits can also be studied using this dataset. After the release of the genotyped and imputed data, WES and WGS sequencing of the samples have been initiated. Part of the exome data, around 50,000 exomes, have already been released and about 200,000 exomes should be expected by the end of 2020. These data foreshadow future key findings in genomics, a better understanding of molecular and phenotypic interactions and probably an improvement of the translation of those findings in the clinic.

Data sharing

With the increasing number of genomics studies, public repositories, like the Database of Genotypes And Phenotypes (dbGAP), the European-Genome Phenome Archive (EGA) or Gene Expression Omnibus (GEO), have been established to store petabytes of genomics data that can be accessed by the research community. In addition, large projects, like the TCGA and ICGC, have worked on solutions to improve data storage and accessibility. One of the goals of those projects was to promote open-access data and the development of tools to foster the reuse of the data by the research community [51, 53]. In 2010, the TCGA provided the data in open access for the first

time [60] and updated and extended the content of the open access data over the years. In 2016, the Genomic Data Common (GDC) was launched by the National Cancer Institute (NCI) to store all the TCGA data [61]. For each omics, the data are categorized by levels: low-level data (raw and unnormalized data) that generally enable individuals re-identification are under controlled access, while higher-level data (processed data, clinical data) that do not permit re-identifiability are available without any requirement. In addition to providing the data storage, the GDC also aimed at harmonizing and sharing the bioinformatics pipelines used to process the data [61, 62]. The processed data resulting from the PanCancer Atlas papers are also available via the NIH GDC website [63] and allow researchers to explore broader genomic features like immune variables [64] or biological pathway measures [65]. Also, cloud computing solutions have been developed to facilitate the analyses of large public genomic datasets while avoiding the download and duplication of the data. The TCGA and ICGC data are available and can be analyzed on the cloud, for example via the Cancer Genomics Cloud (CGC) [66] or the ISB Cancer Genomics Cloud (ISB-CGC) [67]. Also, the ICGC consortium, to process the PCAWG data, has developed a computational tool, Butler, which simplifies genomic analyses that have to be run on clouds environments (academic or commercial) [68].

In the past decades, the development of genomics technologies and the implementation of large consortia have enabled to characterize human cancers on the molecular level. The understanding of cancer causes and the biological mechanisms underlying tumor development has been improved. Also, due to the identification of correlations between molecular events and patient's prognosis and response to treatments, molecular studies have impacted the way that tumors are classified and managed in the clinic.

1.3 The example of lung cancer

1.3.1 Lung cancer subtypes and etiology

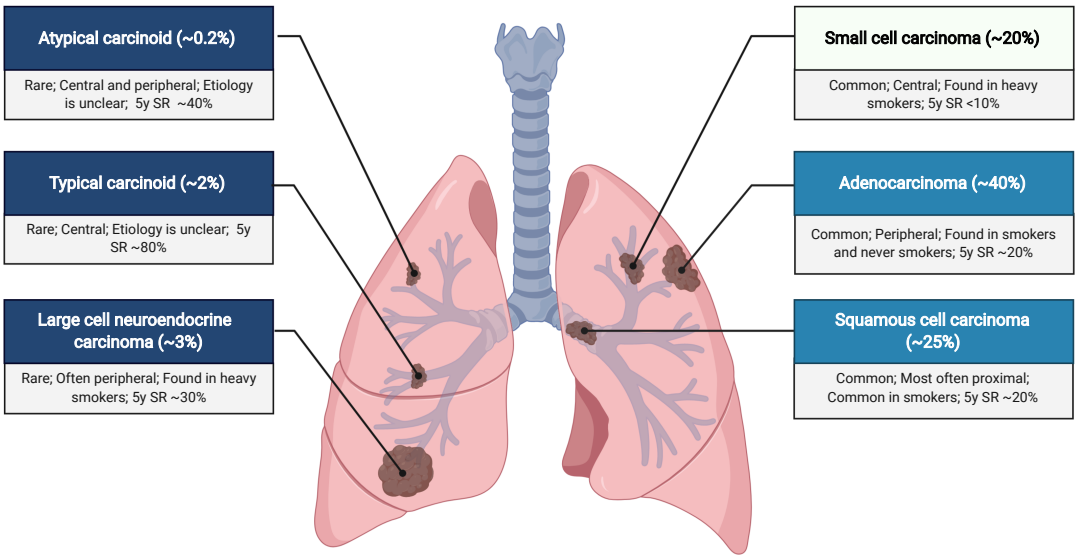


FIGURE 1.9: **Lung cancer subtypes.** Each lung cancer type occurs at different frequencies as well as at distinct locations in the lung (from proximal to distal locations). Each box on the figure is associated to one cancer type and provides their characteristics (frequency, localisation, etiology and overall 5-year survival rate (5y SR)) [69, 70, 71, 72]. Figure created with BioRender.com

As mentioned at the beginning of the dissertation, lung cancer is one of the most common and deadliest cancer worldwide. Several subtypes of lung cancers have been identified (Figure 1.9). The most common lung cancers are usually divided into two groups: the Small Cell Lung Cancer (SCLC) and the Non Small Cell Lung Cancer (NSCLC) samples, representing respectively around 20 and 75% of the lung cancers [73]. The second group is further separated into two main subgroups: the Lung Adenocarcinomas (LUAD) and the Lung Squamous Cell Carcinomas (LUSC). Also, rarer forms of lung cancer exist. Multiple lung cancer subtypes, including such rarer cancers, were grouped in one category named the lung neuroendocrine tumors by the World Health Organization (WHO) 2015 classification [74]. This group comprises the pulmonary carcinoids, including the typical and atypical carcinoids,

Large Cell Neuroendocrine Carcinoma (LCNEC) as well as the previously mentioned SCLC tumors. Each lung cancer type can be distinguished by different etiologies, histopathological characteristics, molecular profiles and clinical outcomes (See Figure 1.9).

The strongest risk factor for lung cancer is smoking. Indeed, SCLCs and LCNECs are frequently found in heavy smokers. Smoking is also a major risk factor for LUAD and LUSC cancers [75]. However, lung cancer can also develop in nonsmokers. In particular, the LUAD category corresponds to the lung cancer type most commonly found in never smokers. Although the etiology of the pulmonary carcinoids is not clear, the majority of these tumors are found in nonsmokers [70]. In addition, around only 15% of smokers develop lung cancer suggesting than other factors mediate lung cancer risk. Such factors include indoor pollution from cooking fumes, radon, and occupational exposures like those from smelting heavy metals or asbestos exposure [76].

1.3.2 Lung cancer susceptibility

While diverse exposures have been identified as lung cancer risk factors, genetics is also contributing to the disease risk. In line with this hypothesis, it has been shown that having a family history of lung cancer confers a 2.5 fold lung cancer risk increase [77]. Further evidence of lung cancer germline susceptibility has been revealed by GWAS studies, with the identification of common variations associated with lung cancer. Genes involved in nicotine addiction (*CHRNA* genes), telomere activities (*TERT*) as well as genes related to the DNA repair and cell-cycle pathways (*e.g. Check2*, *RAD52* or *CDKN2A*) have been identified [78]. Also, some lung cancer associated variants were identified as related to the propensity to smoke [79, 80] and genetic correlations between lung cancer and smoking traits, like smoking initiation, smoking cessation or smoking intensity have been described [80]. Such observations provided evidence that susceptibility variants could influence lung cancer risk through environmental exposures. Hence, GWAS studies have enabled to gain insights on lung cancer etiology as well as on the biological pathways involved in the disease. However, the variants identified so far do not account for most of the heritability of lung cancer, estimated at 18% and remaining today largely unexplained [80].

1.3.3 Lung cancer molecular profiling

In the past decades, molecular profiles of human tumors, including lung tumors, have also been explored thanks to the development of NGS studies. Such studies have, for example, established that lung cancers are among the cancer types with the highest mutational burden (total number of mutations for a given part of DNA) [81]. As mentioned in Section 1.1, in smoking-related cancers, those mutations revealed a signature associated with tobacco consumption. Among the Catalogue Of Somatic Mutations In Cancer (COSMIC) signatures identified by Alexandrov *et al.* [28, 32], the smoking signature corresponds to the Signature 4 (COSMIC version 2) and SBS 4 (COSMIC version 3). Those signatures are the results of DNA damages caused mainly by benzo[α]pyrene, which is a mutagenic compound found in tobacco smoke and whose effects on DNA has been shown in experimental mutagenesis studies [29]. Even though smoking does heavily impact the lung tissue, it has been shown that quitting smoking can restore the damaged tissue [82].

In addition, molecular analyses of lung tumors have identified cancer driver genes in the different cancer types. Among those genes, the *Epidermal Growth Factor Receptor (EGFR)* gene, which is part of the protein kinase family currently known to be mutated in around 15% of the LUAD samples [83], has been related to therapeutic response in 2004 [73]. Indeed LUAD samples, carrying activating mutations in the *EGFR* gene, are responsive to tyrosine kinase inhibitor therapy and have an improved survival in comparison to other cancer patients treated with chemotherapy. Such molecular studies largely influenced the way that lung tumors are classified by leading to the sub-classification of NSCLC. Guidelines were published in 2013 to include molecular testing, mainly based on *EGFR* and *ALK* alterations testing, in the clinical practice for the NSCLC patients. In 2018, those guidelines were updated and new alterations, like rearrangements in the tyrosine kinase *ROS1*, are now recommended for molecular testing [84]. In 2012 and 2014, the TCGA marker papers on the two lung cancer cohorts (LUAD and LUSC) were published. The authors expanded the molecular profiling of these tumors and hence the list of drivers genes, improving the understanding of the biological mechanisms involved and providing new opportunities for patients management [85, 83]. Those studies also explored the transcriptomic, methylation and proteomic data from the lung tumors. Based on their expression profiles, the LUAD tumors, were divided into subtypes that could help to refine those tumors classification [83].

The identification of driver genes in lung cancer has also led to the proposal of molecular targets for early detection. The molecular profiling of SCLCs is an example of such an application. SCLCs are characterized by universal inactivation of

both *RB1* and *TP53* genes [86, 87, 88]. In 2016, Fernandez-Cuesta *et al.* analyzed circulating tumor DNA (ctDNA), which are fragments of tumor DNA released in the bloodstream that can be used as molecular biomarkers, in SCLCs. They showed that *TP53* mutations were detectable in the ctDNA of the SCLC cases [89]. ctDNA applications are viable for multiple cancer types. In 2018, Cohen *et al.* described a blood test called CancerSEEK, detecting proteins and mutations in cell-free DNA for the early detection of eight different cancer types, including lung cancer [90]. Such tests face though sensitivity issues due to the low abundance of mutated DNA in body fluids, hence adapted bioinformatics tools are needed. I contributed to the optimization of such tool, Needlestack, a highly sensitive multi-sample variant caller [91].

Even though rare forms of lung cancers are less explored than the common lung cancers, recent molecular studies have started to characterize the lung neuroendocrine tumors as well [92, 93, 94, 95]. Those studies have revealed that, on top of their histopathological differences, the lung neuroendocrine neoplasms were also distinct molecular entities [88]. Low mutational burden has been observed in the atypical and typical pulmonary carcinoids in contrast to the highly mutated LCNECs and SCLCs [70]. Also, the transcriptomic profiling of those tumors has been investigated. These analyses identified molecular subgroups in different cancer types, revealing the molecular heterogeneity in those tumors [93, 96]. The work described in chapters 2 and 3 of this thesis contributed to the molecular characterization of the lung neuroendocrine tumors.

The discoveries described in this section were enabled thanks to the large amount of data generated during the era of genomics (See Section 1.2). However, the analyses of these data have raised multiple challenges that required the use and development of specific computational methods. The next section intends to describe those aspects.

1.4 Interpreting high dimensional data

The evolution of genotyping and sequencing technologies led to the generation of high dimensional datasets. In Section 1.2, we have seen for example that arrays can interrogate thousands to millions of positions across the genome and that sequencing techniques can provide the entire genome sequence or the expression levels of thousands of genes. While the amount of information unveiled by these methods is

colossal, it can also bring about several challenges and adapted computational methods are required to analyze and interpret the data. The issues resulting from high dimensionality are associated to what is called the curse of dimensionality, firstly introduced by Bellman in 1961 and stipulating that the number of samples needed to interpret high dimensional data analyses appropriately increases exponentially with the number of dimensions [97]. In omics datasets, even though large cohorts have been implemented (see section 1.2), the number of variables (also known as features), p, to analyze can be largely superior to the number of samples, n, included in the study. This introduces the $n << p$ problem, which leads to multiple issues. Firstly, usual statistical models like regression models need to be adapted since they require $p < n$. There is also a substantial amount of noise in the generated data that can mask the true signal in the data, *i.e.* not all the measured features are of interest [98, 99]. In addition, when the number of dimensions increases, the data points can occupy a more voluminous space and a larger proportion of this space will be empty, we say that the data are sparse (See Figure 1.10) [97]. High data sparsity influences basic properties to which we are used to in two or three dimensions like distances. In high dimensions, distances between points increase and all points seem at the same distance from each other [99, 97]. Also, the higher the dimensions, the lower the correlations between the features will be. For those reasons, it is thus statistically more difficult to identify groups of points with similar characteristics compared with random events, as such larger sample sizes are required to distinguish meaningful relationships. Another issue resulting from high dimensionality is multi-collinearity. Since the number of features is high, the information they carry can be correlated and become redundant; some variables might be defined as a linear combination of others which makes the data interpretation more difficult [97]. Finally, the nature of omics datasets complicates the visualization of the data. In this section, we will discuss in a first instance different strategies to explore such complex datasets and secondly focus on methods that attempt to diminish the problem of the curse of dimensionality: the dimensionality reduction methods.

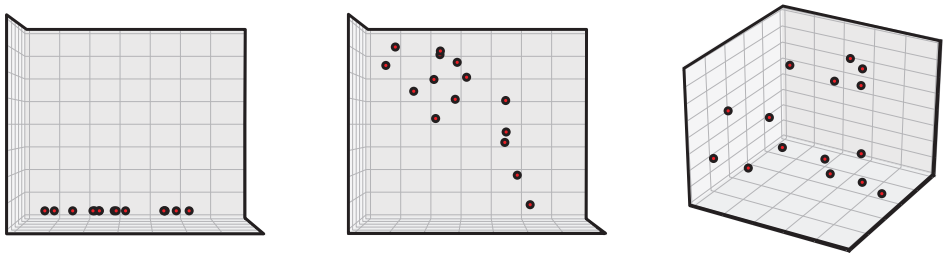


FIGURE 1.10: **Illustration of data sparsity.** Figure from [99]. The figure represents how the data occupy the available space when going from a one-dimensional space to two and three-dimensional spaces (from left to right panels).

1.4.1 Supervised and unsupervised methods

Different approaches exist to analyze high dimensional data like omics data. In the case where specific biological hypotheses need to be tested, confirmatory data analyses based on inference models can be used. It can also happen that there are no predefined hypotheses and that the goal is to "let the data talk", in that case, exploratory data analyses (EDA) will be more adapted [100]. A broad panel of statistical methods exists to assist both approaches. Among them, a large proportion can be grouped in the popular category of machine learning methods. The term machine learning (ML) was used for the first time by Arthur Samuel around 1950 and defined a group of computer algorithms able to learn without being explicitly programmed to learn. Depending on the definition of learning, different classes of ML methods have been established. In 1997, Tom Mitchell proposed a formal definition of algorithms learning saying that "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [101]. This definition matches a class of ML methods, the supervised learning methods, used for classification and regression tasks. A common example is the identification of spam emails, where labelling emails in the spam or non-spam categories would be the task T, learning from a set of labelled emails would be the experience, and the proportion of correctly classified emails would be the performance measure P. However, ML algorithms that simply learn from the input dataset without predefined ground truth (labelled data) also exist and are part of the unsupervised ML methods. Those methods learn underlying structures in the data; hence algorithms like clustering or dimensionality reduction methods such as Principal Component Analysis (PCA), which was developed even before ML, are often included in the unsupervised learning category. In the next paragraphs, both supervised and unsupervised learning are described (See Figure 1.11).

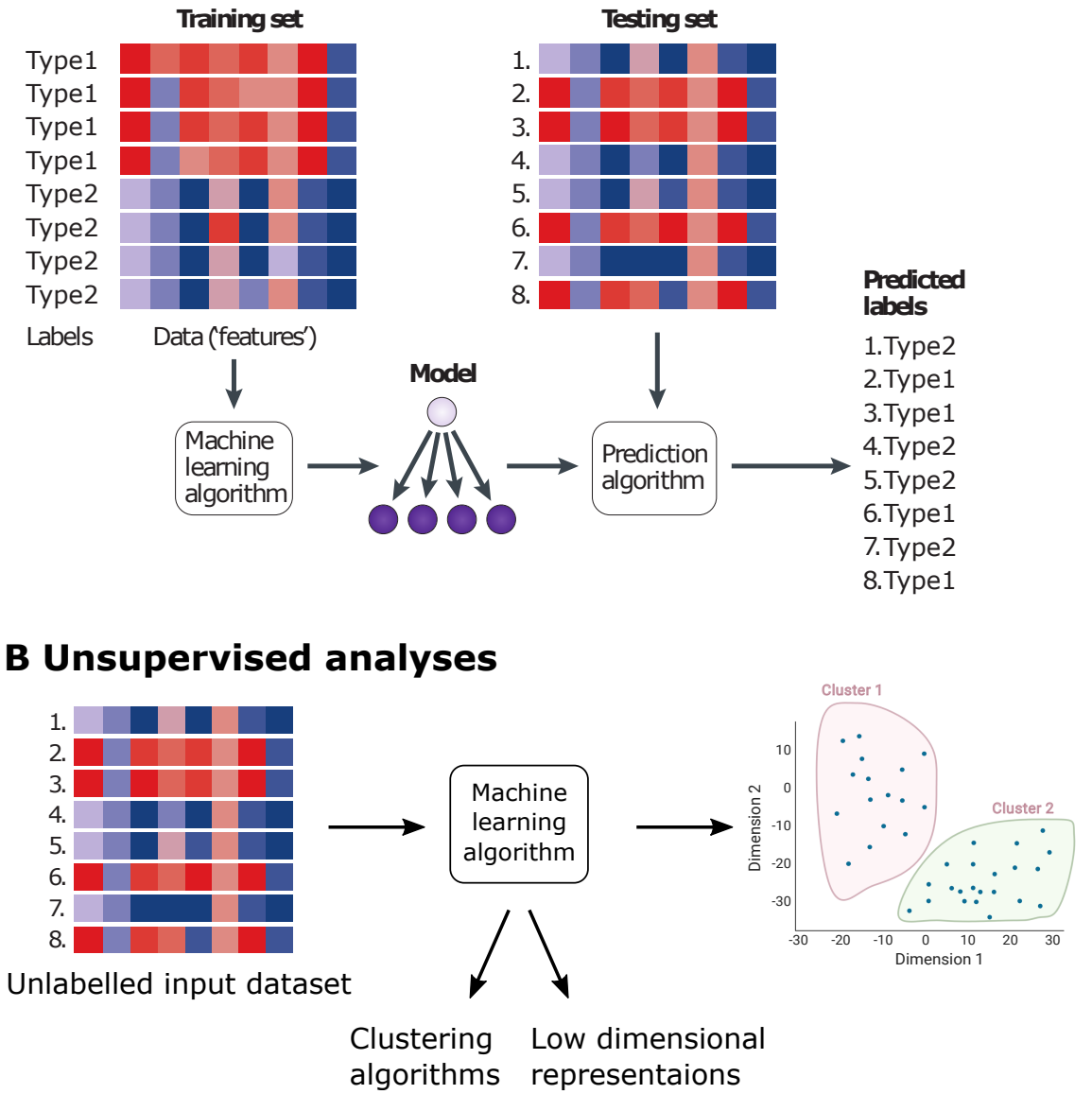


FIGURE 1.11: **Machine learning methods: supervised vs non-supervised methods.** A) Supervised methods: a model is trained on several variables, features, to recognize predefined labels. The trained model is then applied to an unlabelled dataset for prediction purposes. B) Unsupervised methods: a model learns structures underlying a dataset that has not been labelled. Those methods are divided into two main categories: clustering methods to identify subgroups of samples and dimensionality reduction methods to explore the data in lower dimensions and highlight specific structures. Figure adapted from [102].

Supervised analyses

The goal of supervised methods is to predict the value of an outcome based on a set of features given as inputs. Depending on the type of outcome, supervised

analyses can be further divided into two main categories: classification or regression problems. In classification problems, the outcome is categorical, *e.g.* a binary variable distinguishing a diseased or healthy status or a multi-classes variable like cancer subtypes. In regression problems, the objective is to predict a continuous variable. Note that some regression models, like logistic regressions, where the outcome variable is discrete, can be used though to perform classification. The main steps of supervised analyses consist in: i) defining the labels of each sample in the dataset, ii) train the model to classify the samples in the correct category, and iii) use the generated model on a dataset containing independent and unknown instances (Figure 1.11A). Several types of supervised methods exist and have to be chosen with regard to the nature of the data. The simplest supervised models are regression models. While the most common regression algorithms model linear relationships, other methods like Support Vector Machines (SVM) or neural networks can adapt to non-linear data. Another parameter that determines the type of methods to use is the data type; some methods deal only with numerical features while others like decision trees are more flexible. Figure 1.12 describes a method based on decision trees, the random forest algorithm.

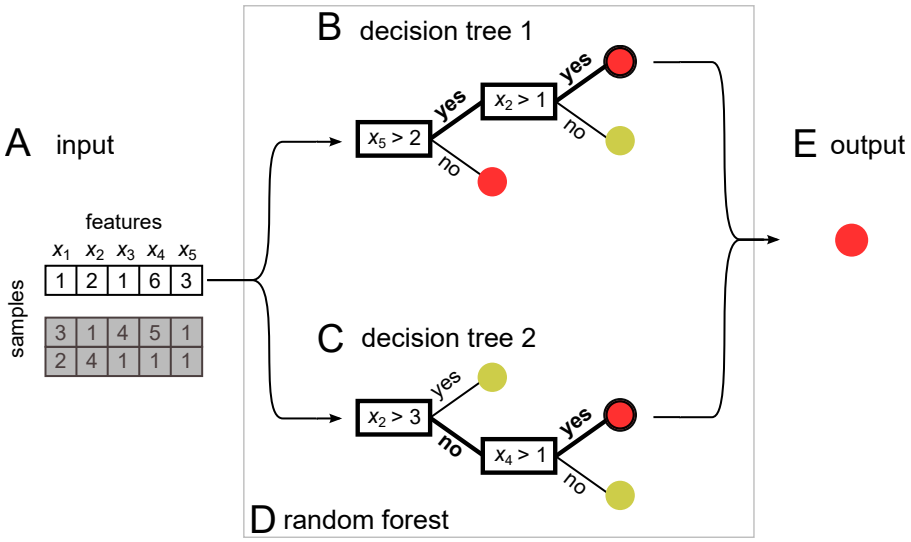


FIGURE 1.12: **The random forest method.** Figure from [103]. A labelled dataset (A) is taken as input and processed by multiple decision trees (B and C) built using random selections of features and samples. The decision trees form a random forest (D). Each tree classifies the input samples and the votes given by the different trees are then combined to provide the final predictions. The label with the most votes being chosen (here red label).

Regardless of the method used, the model and its results have to generalize to other datasets. In order to assess generalizability, the ML algorithm has to be trained

on a training dataset, and a testing dataset containing independent samples has to be used to validate the results. Two main errors underlying the generalization issue exist: bias and variance [98]. The first scenario occurs when the model is underfitting the data, *i.e.* the model has a poor performance even on the training data for example because of a model that is not complex enough (See Figure 1.13 left panel). When the model is underfitting the data, it is as well unable to generalize to other datasets. In the second case, when the number of features is too large or the number of samples small, the chances to encounter features that can perfectly discriminate two output categories or perfectly predict an outcome increase. The model, in that case, performs correctly on the training dataset but fails to generalize to other datasets and is qualified as high variance model. Such performance discrepancy indicates that the model overfits (See Figure 1.13 right panel). Note that in high dimensional data, overfitting and data sparsity, resulting from the $n << p$ problem mentioned at the beginning of this section, can be linked. Indeed, in such data, since the number of samples in the training dataset is fixed and limited, the entire input space is not covered. Thus the machine learning algorithm has not faced all possible configurations during the learning phase and the ability of the model to generalize can be diminished.

FIGURE 1.13: **High bias and high variance models.** Created with BioRender.com.

One method that can be used to overcome overfitting is cross-validation. The method consists in randomly splitting the dataset in k folds and iteratively training the model on $k-1$ folds while reserving the remaining kth fold for testing (See Figure 1.14). The overall performance of the model can be assessed by averaging the performances in the testing folds from each iteration. As a result, while none of the samples is used simultaneously in the training and testing group, the entire dataset is used for training as well as is used in the testing phase. Hence, cross-validation can also be beneficial in studies with low sample sizes. One extreme case of cross-validation is the leave-one-out analysis, where $k = 1$. Each sample is set aside from the training set and predicted at each iteration.

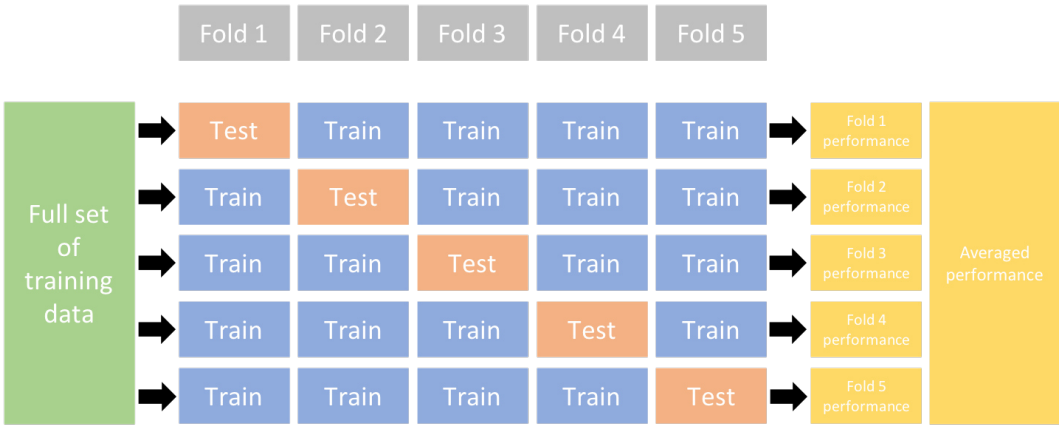


FIGURE 1.14: ***K*-fold cross-validation.** Figure from [104]. The figure illustrates 5-fold cross-validation. Five rounds are thus represented. In each of them, 4 folds are used to train the model and the model is tested on the remaining fold. The performances resulting from the test phase in each round are then averaged to estimate the overall performance of the model and its ability to generalize.

In addition, to find a compromise between bias and variance, parameter tuning and algorithm optimization might be required. Note that a third dataset, referred to as the validation dataset, can be introduced for the optimization step. In this setting, multiple models (*e.g.* one algorithm with different sets of parameters or different algorithms) learn on the training set, and their performances are evaluated on the validation dataset. The model with the best performance can then be applied on the testing dataset.

Unsupervised analyses

Unsupervised algorithms are hypothesis-free methods and can be associated to exploratory analyses [105]. The goal of such methods is usually to identify and extract useful properties of the data [106]. In contrast to the supervised methods, each element of the dataset is not labelled, no predefined groups are given to the algorithms. Thus, it is not possible to compare the algorithm output with a predefined truth and the data do not need to be split in training and testing datasets (Figure 1.11B). Since there is thus no feedback on the performance of the unsupervised model, often the validation of the results is required.

As for the supervised analyses, there are several unsupervised algorithms. A commonly used category of unsupervised methods that can unveil structure in the data is the group of clustering algorithms (*e.g. k*-means clustering, hierarchical clustering, density-based clustering). Those methods aim at grouping elements together

based on common patterns observed in the set of features. In the field of cancer, clustering algorithms can be used, for example, to identify new subtypes of cancers based on molecular data. The second most commonly used unsupervised method is the group of dimensionality reduction methods. In the next paragraph, more details about such methods are provided.

1.4.2 Dimensionality reduction methods

The goal of dimensionality reduction (DR) methods is to transform a high dimensional dataset into a low dimensional representation of the data while preserving as much as possible its initial structure. More specifically, if three clusters exist in the studied dataset, a lower dimensional representation of the same data should also reveal the initial three clusters. DR methods are part of the feature extraction techniques which aim at finding latent structures in the data. Those methods allow to summarize and transform a large number of features in a smaller number of variables, which mitigates the curse of dimensionality and is valuable for data visualization. Note that these methods are different from feature selection methods, which make a selection of the most important features in the initial dataset [107]. Mainly two families of DR methods exist: matrix factorization methods (*e.g.* PCA, PLS, ICA, NMF) or neighbour graphs approaches (*e.g.* t-SNE and UMAP).

Matrix factorization methods examples

Omics datasets, after pre-processing, often result in data matrices. For example, in the case of RNA-Seq, after aligning the reads to a reference genome (See Figure 1.8), reads counting is performed and generates a matrix in which rows represent the genes (the features) and columns the read counts for each sample (the observations). Matrix factorization consists in decomposing an initial matrix in two smaller matrices (Figure 1.15). This decomposition leads to the generation of new variables, in smaller numbers.

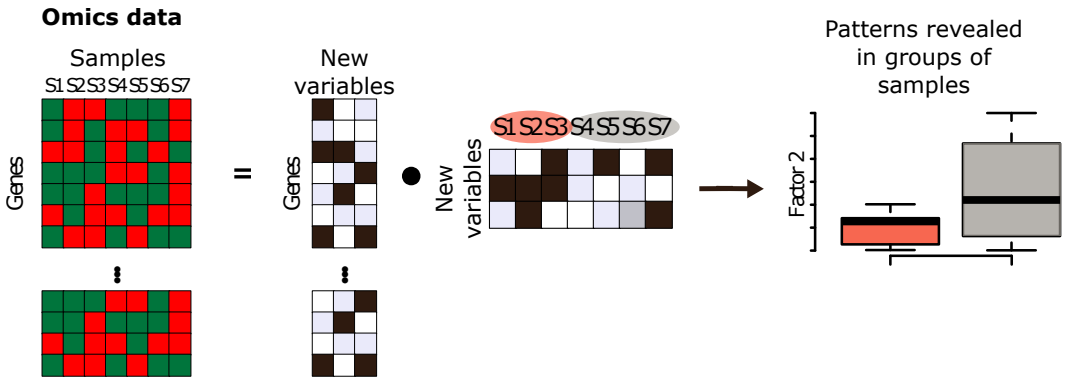


FIGURE 1.15: **Matrix factorization methods**. The input matrix is decomposed, under specific constraints, in two smaller matrices defined by new variables that can be used to reveal structures and patterns in the data.

A classical matrix factorization method is Principal Component Analysis (PCA). The goal of PCA is to project the data to a lower dimensional space while maximizing the variance in the data within this lower dimensional space. In PCA, the new variables correspond to a linear combination of the initial features. The matrix factorization results in the loading and score matrices. In the first matrix, the columns correspond to the new variables, called principal components and the rows indicate the contribution of each feature to the latent variables. The principal components are orthogonal; they correspond to the directions of maximal variance and are ranked by the importance of variance explained, *i.e.* the first principal component captures most of the variation in the dataset. The second matrix contains the coordinates of the samples in the projected space. While PCA maximizes the variance in the data, similar methods use other criteria. For example, Independent Component Analysis (ICA), which is a method attempting at disentangling independent signals that are linearly mixed, maximizes the independence between the new variables. Other methods have in addition specific constraints [108]. Non-negative Matrix factorization (NMF), for example, enforces the decomposed matrices to be positive; this method has enabled the extraction of *de novo* mutational signatures from whole genome sequencing data [109]. One limitation of those methods is that they are linear models. In the following paragraphs, two non-linear methods based on neighbour graphs are presented.

Neighbor graphs methods examples

The principle of DR methods based on neighbor graphs models is to use neighbors distances and similarities to represent the structure of the data in high dimensions and then to embed this representation in a lower dimensional space.

A method called t-Distributed Stochastic Neighbor Embedding (t-SNE) [110] has been widely used in the past years to perform DR. The t-SNE method can be seen as a neighbor graph based algorithm [111] in a sense that similarity scores based on Euclidean distances between neighbors are computed to embed the high dimensional structure in a two-dimensional space. Samples positions in the two-dimensional space are randomly initialized and are then moved iteratively so that the pair-wise samples similarities match the ones in the original space. t-SNE has limitations though. Firstly, the method can be computationally intensive when applied to huge datasets. Also, the interpretation of the t-SNE representation must be performed with caution. Indeed, the method retains local structures but has limited ability to maintain global structure [111].

Recently, a novel method called Uniform Manifold Approximation and Projection (UMAP) [111] was developed and is more and more replacing the t-SNE method. UMAP is based on topological theory. The algorithm builds what is called a simplicial complex which is a representation of the data as a weighted graph (See Figure 1.16), the weights corresponding to the likelihood that there is a connection between two points [112, 113].

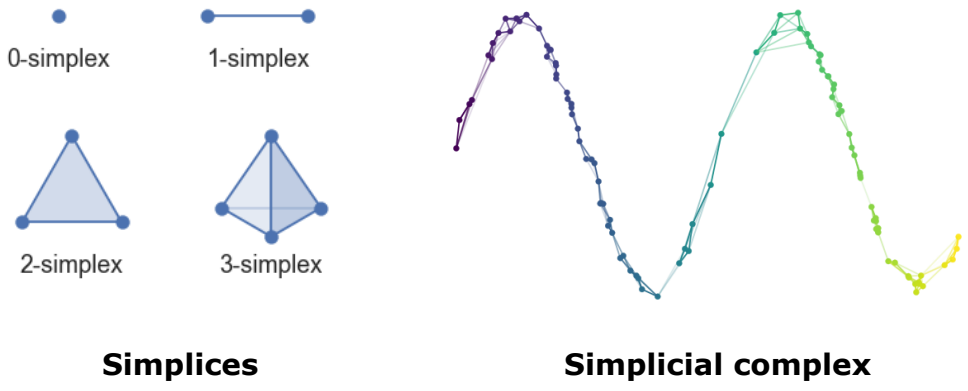


FIGURE 1.16: **UMAP topological representation**. A) The building blocks of a simplicial complex, the simplices. B) An example of a simplicial complex. Figures from [112].

As mentioned at the beginning of Section 1.4, in high dimensional spaces data sparsity increases. To connect all the points in the simplicial complex, UMAP varies the radius in which the search of neighbors is performed by fixing the number of

neighbors to consider around each point [112, 113]. This number of neighbors influences how the data structure is preserved, low and high values favoring local and global structures, respectively. Once the graphical representation of the high dimensional data is constructed, a low dimensional representation of the data is optimized so that it is as close as possible to the high dimensional representation. One of the advantages of UMAP over t-SNE is that the method better maintains the global structure of the data. Also, UMAP is computationally more efficient [111]. Note that UMAP can be applied on a lower dimensional dataset resulting, for example, from a DR method like PCA.

1.4.3 Multi-omics data integration

The methods previously described consider as input a single dataset. DR methods processing multiple matrices also exist and can be used to integrate multi-omics datasets. Such integration raises, though, multiple challenges. Firstly, the data to integrate are heterogeneous. The nature of the collected data is different, hence their statistical properties can vary. Also, it can happen that all the omics datasets are not available for each sample included in the analysis for technical reasons or due to quality issues. Hence, distinct patterns of missing data can occur in each omic dataset. Besides, integrating multiple datasets amplifies the curse of dimensionality issues already encountered in each dataset individually.

In 2018, a method called Multi-Omics Factor Analysis (MOFA) was developed to integrate multi-omics data while considering the previously mentioned challenges [114]. MOFA is an unsupervised analysis based on matrix factorization (See Section 1.4), and can be seen as an extension of PCA to multi-omics data, called modalities or also views. It is a factor analysis method which reduces the dimensions of the data to a smaller number of unobserved factors, called the latent factors. These factors differ from the PCA components. The latter are linear combinations of the initial features, while in factor analyses the initial features are expressed as linear combinations of the latent factors, plus a residual noise term. To enable multi-omics data (modalities) integration, MOFA supports different noise models depending on the nature of the data (continuous, counts or binary data). Based on this model, MOFA identifies different sources of variations across multiple omics data. MOFA presents though several limitations. The model does not capture non-linear relationships and assumes features independence [114]. Also, additional features accounting for samples structure, such as groups of samples, batches or samples conditions, were not available in the initial version of MOFA but have been recently introduced in a second version, MOFA+ [115]. In this framework, the MOFA dimensionality reduction

is performed with regards to additional samples information (*e.g.* batch or cluster information) to identify sources of variations shared between groups or exclusive to one of them.

Other integrative methods can take into consideration samples structure. For example, the Partial Least Squares (PLS) method, which is a matrix factorization method, attempts to relate two matrices: a response matrix and a matrix gathering explanatory variables. The advantage of this method is that it ensures that the new variables resulting from the dimensionality reduction explain the response data. In that sense, the PLS method can be considered as a supervised DR framework. While PCA maximizes the variance of the components, PLS maximizes the covariance between the latent components of the response and explanatory datasets [116, 107]. When the response data is a categorical variable, a variant of PLS called PLS discriminant analysis (PLS-DA) can be used to perform classification tasks, *e.g.* samples groups prediction. In 2017, Lê Cao team published the mixOmics framework implementing multivariate analyses tools, including the PLS methods previously described [117]. The mixOmics tools also include the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) method, which is a multivariate dimension reduction method that can be used for supervised multi-omics data integration [118]. DIABLO maximizes the correlation between the features of the different omics datasets, one of this dataset corresponding to the labelled samples. Hence, the method extracts what the authors call multi-omics signatures that are discriminant and can be used for prediction in a supervised framework.

1.5 Axes of the thesis

In the previous section, we discussed different methods available to analyze high dimensional omics datasets and overcome some of the challenges related to the curse of dimensionality. Another challenging aspect of omics data is their biological complexity. As described in Section 1.1, multiple biological layers (*e.g.* genome, transcriptome, methylome, *exposome*) interact. The work presented in this thesis highlights how integrative approaches can improve the understanding of such complex systems by relying on the computational methods described previously and using lung cancer as an example.

Firstly, while single-omics approaches explain a substantial amount of mechanisms involved in cancer, it is difficult to capture all the complexity of the disease using each omic layer individually [119]. The integration of multi-omics data could

thus expand our understanding of cancer. Chapter 2 of the thesis describes the multi-omics characterization of lung neuroendocrine neoplasms. As mentioned in Section 1.3, those rare tumors have not been comprehensively characterized, especially the pulmonary carcinoids. While a low mutational burden has been observed in those tumors, more insights on their carcinogenesis might be provided by other omics data like expression and methylation. Machine learning methods were applied on RNA-Sequencing and methylation arrays data to reveal differences and similarities between the lung neuroendocrine neoplasms (lung NENs) (or LNEN) cancer types.

Secondly, the increase of large genomics initiatives has enabled to perform analyses contrasting the molecular profiles of distinct tumor types. Such studies implying the concatenation of datasets from different studies raise challenges, including data harmonization and interpretation. In chapter 3, we integrated six transcriptomic datasets from Lung Neuroendocrine Neoplasm (LNEN) tumors in order to produce a pan-LNEN molecular map. This map and especially the underlying data are intended to be reused and integrated with future similar datasets. For that purpose, the pre-processing and the quality controls performed on the data were described precisely, and additional resources promoting reproducibility and data reuse were provided.

The two first chapters focus on somatic molecular characterization of lung cancers. In the past decades, lung cancer susceptibility has also been explored mostly by GWAS studies which revealed multiple variants across the genome. However, the identification of causal genes involved in lung cancer susceptibility has raised challenges. While adding information from other biological layers (*e.g.* expression data) has been proposed to identify causal genes, investigating the germline and somatic interplay could also bring new insights on lung cancer oncogenesis. In the final chapter, we integrate germline and somatic data from lung adenocarcinomas and lung squamous cell carcinomas in order to explore the association between lung cancer susceptibility variants and mutational burden.

Chapter 2

Somatic molecular characterization of lung neuroendocrine neoplasms using multi-omics data

2.1 Context

Today cancer is considered as a very heterogeneous disease, each cancer is different. Under the microscope, pathologists can distinguish cancers from different tissues but also subtypes originating from the same primary site. Characterizing the molecular landscape of tumors has confirmed these observations and can be leveraged to understand this heterogeneity and its consequences for the patient. Firstly, even if tumor classification is still mostly based on histopathological criteria, molecular studies have introduced ample changes in the way that tumors are classified. Lung cancer is a good example to illustrate how molecular profiles have assisted this shift in cancer diagnosis. While a few decades ago, lung cancers were stratified in only two categories, SCLCs and NSCLCs, it is now clear that there is a need for a more precise classification. As mentioned in the introduction section 1.3, molecular profiling of those tumors has identified recurrent alterations in specific subtypes, like the *EGFR* mutations in LUAD samples, influencing the patient's prognosis and response to different intervention therapies [120, 73]. Molecular studies have also improved the way cancers are diagnosed by identifying new targets for early detection. The use of ctDNA methods has, for example, been developed for some cancers early detection [90, 121]. These non-invasive methods aim at detecting alterations found in tumors and released in body fluids to diagnose cancer at the early stages of the disease. In addition, genomics datasets have enabled to provide a molecular-based taxonomy of cancer [122]. Cancer types have been stratified in molecular subgroups that can be distinguished by different biological pathways as well as different prognosis. The molecular characteristics identified can thus be used as biomarkers at

diagnosis to inform the tumor classification and the patient's prognosis. In this context, gene expression profiling, which is a technique already used in clinical practice for example for breast cancer [123] is an illustration of the potential clinical translation of multi-omics studies. Finally, an increasing number of studies has highlighted the importance of integrating multi-omics data for molecular characterization. Indeed, driver events can impact multiple omic layers. Depending on the type of alteration, some layers might be more adapted than others to detect alterations [119]. Also, in cancers with low mutation rate, exploring the transcriptome and epigenome of tumors as well as their interactions can bring new insights on their carcinogenesis [124].

In the last years, comprehensive somatic molecular characterizations of various tumors have been performed, in part thanks to multiple initiatives like the TCGA or ICGC. However, those projects mainly focus on the characterization of common subtypes. Hence more studies are still required for rarer cancers that collectively represent around 25 to 30% of cancer diagnoses and 25% of cancer deaths [125]. Indeed, the lower incidence of those cancers is a major limitation for such studies. In the context of rare cancers, identifying new genomic alterations may provide to researchers new targets for the diagnosis, classification and treatment of the patient's tumor. Recently, the Rare Cancer Genomics initiative [126] has been developed by Dr. Lynnette Fernandez-Cuesta and Dr. Matthieu Foll and aims at performing a molecular characterization of rare cancers, including the rare lung neuroendocrine neoplasms. In this chapter, we explored the molecular landscape of the lung neuroendocrine tumors using integrative analyses of multi-omics data.

2.2 Research contribution

2.2.1 Introduction

The lung NENs represent 25% of the lung cancers and are divided into subgroups (Figure 2.1) [88]. On one hand, the atypical carcinoids and typical carcinoids form the group of the rare neuroendocrine tumors (NET) and account for 2% of the lung NENs. On the other hand, LCNECs and SCLCs account for 3% and 20% of the lung NENs respectively and are part of the neuroendocrine carcinomas (NECs). The four types differ at different levels. NECs are high grade carcinomas, have a poor prognosis and require aggressive treatments; while typical and atypical carcinoids are low- and intermediate-grade tumors respectively, show a better prognosis and are eligible for surgical resection [88]. Thus, the proper clinical management of lung

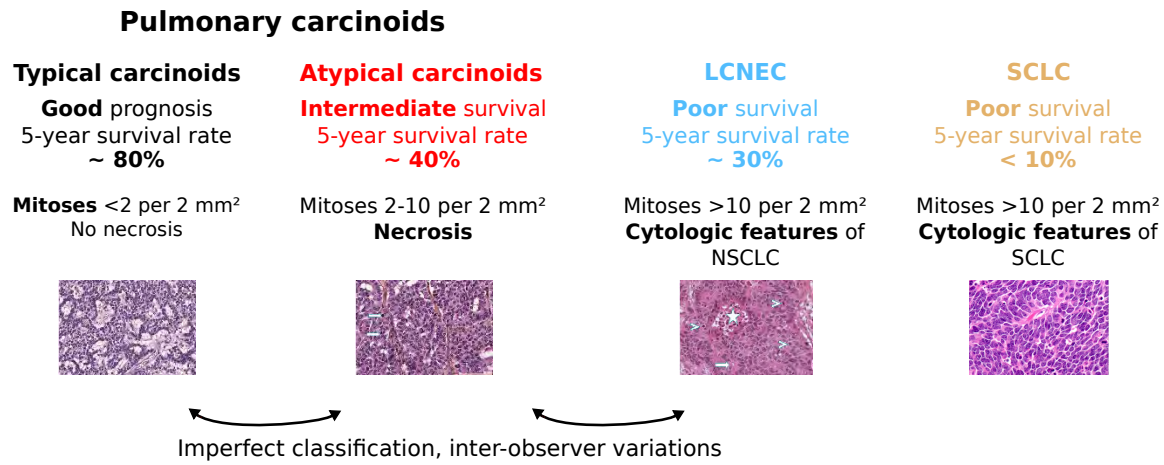


FIGURE 2.1: **The different types of lung neuroendocrine neoplasms**. Four types of lung NENs exist: the typical and atypical carcinoids that form the group of the pulmonary carcinoids, the LCNECs and the SCLCs. The four types have distinct prognosis and are classified mostly based on histopathological criteria.

NENs rely on an accurate classification. Currently, this classification is based on histopathological criteria such as the number of mitosis and necrosis as well as on immunohistochemistry markers [88, 71]. However, those criteria are imperfect and a consensus is often difficult to reach. Indeed, the study of Swarts *et al.* [127] has assessed the reproducibility of pulmonary carcinoids classification by contrasting the diagnosis of five pathologists and showed that only 55% of the cases were unanimously classified [127]. Hence, molecular studies on these tumors could help to identify new biomarkers and improve their diagnosis. In addition, even if the number of molecular studies on these cancer types has increased in the past years [86, 128, 92, 87, 93], their etiology has not yet been clearly determined, the mechanisms underlying their oncogenesis remain unknown and their therapeutic opportunities limited. The lack of markers for these cancers is thus a drawback not only for the proper diagnosis but also for the follow-up of the patients.

In the study presented in this chapter, we performed a molecular characterization of the lung NENs (or LNEN) with a particular focus on the understudied atypical carcinoids. Multi-omics data were integrated using supervised and unsupervised methods to better understand the differences and relations between the LNEN subtypes in order to improve diagnosis and management.

2.2.2 Material and methods

For this study, WES/WGS, RNA sequencing and EPIC 850k methylation array data from 83 lung NENs, including 63 carcinoids and 20 LCNECs, have been generated.

The samples were obtained in the context of the LungNEN network. They were collected based on surgical resection and were reviewed by independent pathologists. These newly produced data have been integrated with previously published data [86, 92, 87, 93], increasing thus the samples size to 257 LNEN samples including 116 carcinoids, 75 LCNECs and 66 SCLCs. Integrative methods, considering the multiple layers of omics data, have been used to perform a molecular characterization of the different subgroups.

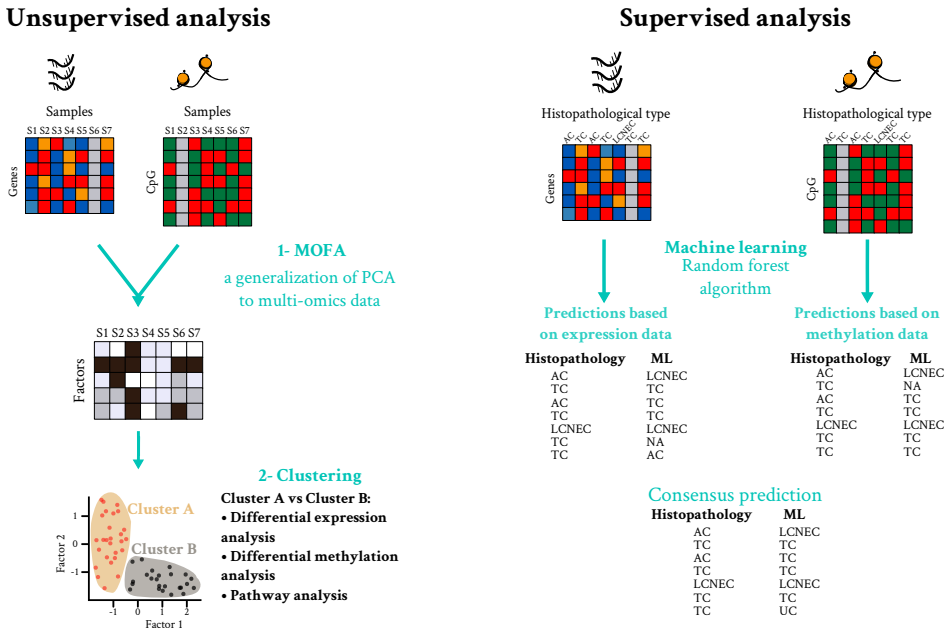


FIGURE 2.2: **Description of the methods used for the LNENs molecular characterization**. A) Unsupervised analysis using MOFA based on expression and methylation data. B) Supervised analysis based on the random forest algorithm applied on expression and methylation data.

Pathologists classified each LNEN sample in this dataset in the different LNEN cancer types. As described previously, this classification is difficult, especially for the pulmonary carcinoids types. The use of a supervised machine learning (ML) model on the available omics data was thus suited to assess if molecular data could predict those cancer types and assist histopathological classification. We applied random forest algorithm, which is a classifier based on decision trees (See introduction section 1.4), to the expression and methylation data in order to distinguish atypical carcinoids, typical carcinoids and LCNEC samples (Figure 2.2). Since both omics data were not available for all samples, the method was applied to expression and methylation data separately to maximize the samples size for further analyses. Also, as described in the general introduction section 1.4, multi-omics data that are high dimensional datasets and are prone to overfitting issues. Indeed, the samples

sizes of the expression and methylation datasets in this study were small in contrast with their respective number of features. For this reason, the leave-one-out method (see introduction section 1.4) was used to classify each sample. The most variable features were selected and normalized on the training set, consisting of all the samples minus the test sample to classify. The model was trained on this training set and used on the test sample to compute the probability of belonging to each of the three histopathological groups. For each case, the group with the highest probability was defined as the ML prediction. However, when the ratio between the two highest probabilities was higher than 1.5, the sample was considered as "unclassified". This category allows identifying samples with an intermediate molecular profile. For each sample, we then compared the expression and methylation-based predictions to reach a consensus. When the two omics layers led to discordant classifications, the sample was classified in the "unclassified" category.

In parallel, an unsupervised analysis has been performed using multi-omics factor analysis (MOFA) (Figure 2.2) [114]. MOFA, as described in the introduction (See general introduction Section 1.4), is a generalization of PCA to multi-omics data. It thus allows to reduce the dimension of the multi-omics datasets by identifying latent factors that unveil multiple sources of variations that are either shared by the different omics data or specific to one layer. As for the supervised approach, MOFA was applied using two layers: the expression and methylation data. The two first latent factors, capturing most of the variance in the datasets, were then used to perform a consensus clustering in order to reveal distinct molecular subgroups. These unsupervised methods were applied on the LCNECs and carcinoids as well as on carcinoids only. The molecular groups identified by the clustering were finally characterized based on differential expression and methylation analyses as well as Gene Set Enrichment Analyses (GSEA). Also, to support the clinical relevance of the molecular groups identified, survival analyses were performed using Cox's proportional hazard models.

2.2.3 Results

A supervised analysis based on random forest was performed to predict the LNEN histopathological categories based on molecular features like expression and methylation levels.

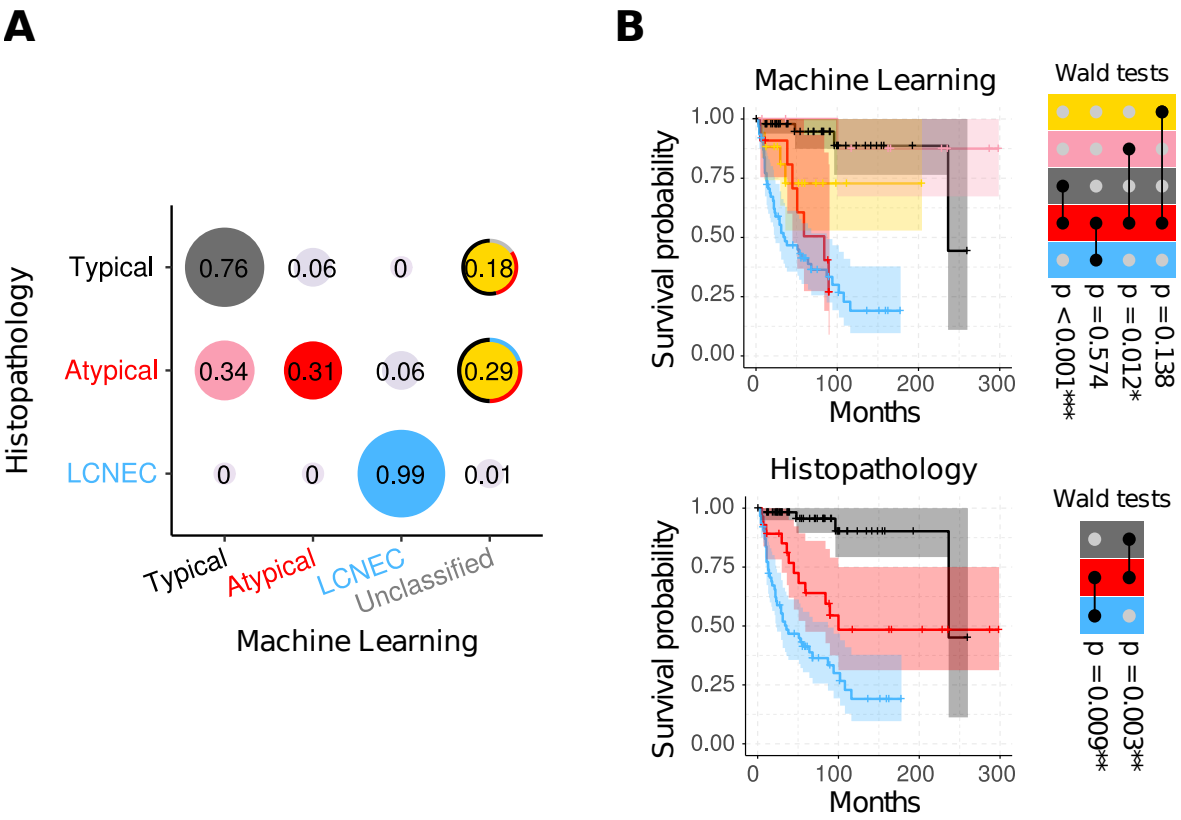


FIGURE 2.3: **Supervised analysis results.** A) Confusion matrix representing the ML classifications, the histopathological categories are represented on the x-axis and the ML predicted classes on the y-axis. B) Comparison of the survival between the ML prediction groups (top panel) and comparison of the survival of the LNEN histological classes (bottom panel).

Figure 2.3A represents the prediction results in a confusion matrix and shows that the classifier accurately distinguished LCNEC samples from the pulmonary carcinoids since 99% of the LCNECs were correctly classified. However, the distinction between atypical carcinoids and typical carcinoids appeared to be more difficult. Only 31% of the atypical carcinoids were classified as atypical, while another third of atypical carcinoids was predicted as typical carcinoids. Also, for 18% and 29% of the typical and atypical carcinoids respectively, the classification algorithm hesitated between two subtypes. These samples, that we labelled "unclassified samples", are representative of samples with intermediate molecular profiles. These results show that the molecular data do not perfectly match the histopathological classification of the lung NENs. Since the histopathological classification was used here to train the model, this outcome indicated that the current classification might not be appropriate and that molecular data could bring additional information to stratify LNEN samples. Based on the supervised analysis results, we defined five distinct groups of ML predictions: i) the atypical carcinoids predicted as atypical

carcinoids, ii) the atypical carcinoids reclassified as typical carcinoids, iii) the typical carcinoids predicted as typical carcinoids, iv) the LCNECs predicted as LCNECs, and v) the unclassified samples. We then compared their overall survival (Figure 2.3 top panel). The survivals of the two groups of atypical carcinoids were significantly different. The atypical carcinoids reclassified as typical carcinoids had a better survival, similar to that of the typical carcinoids group and the atypical carcinoids confirmed as atypical carcinoids had a poor prognosis, similar to that of the LCNEC group (10-year overall survival of 88% and 27% respectively). This observation was in contrast with what is observed when comparing the survival of the histopathological groups (Figure 2.3 bottom panel). The samples diagnosed as atypical by the pathologists showed indeed an intermediate survival in comparison with those diagnosed as LCNEC and typical.

The MOFA analysis based on the pulmonary carcinoids and the LCNECs revealed three clusters: cluster A enriched for typical carcinoids, cluster B enriched for atypical carcinoids and cluster LCNEC mainly composed of LCNEC samples. While each cluster was enriched for one of the histopathological group, atypical and typical carcinoids were not clearly separated, based on their molecular profiles. These results are concordant with the supervised analysis. More specifically, most of the atypical carcinoids clustering in the typical-enriched cluster A were predicted by the ML algorithm as typical carcinoids. The atypical carcinoids confirmed atypical by the ML were part of the cluster B. Also, the intermediate molecular profiles identified based on the supervised approach were borderline samples when considering the unsupervised molecular clusters. In addition to the clusters, the MOFA analysis unveiled a novel sub-group of pulmonary carcinoids, the supra-carcinoids. These samples have the morphological features of the atypical carcinoids but molecular features and survival similar to that of the LCNECs. On the molecular level, those samples were characterized by high expression levels of immune checkpoint inhibitors and Major Histocompatibility Complex (MHC) class I and II genes.

Finally, the MOFA analysis performed only on the carcinoids samples stratified the cluster A in two groups, clusters A1 and A2. GSEA analyses performed on the MOFA latent factors identified the immune system and the retinoid and xenobiotic metabolism as disrupted pathways in the pulmonary carcinoids. Using expression and methylation data, the clusters A1, A2 and B were further characterized on the molecular level. This molecular characterization highlighted potential candidate targets with potential clinical applications. For example, *DLL3*, an inhibitor of the

Notch pathway, which is already considered in clinical trials as a target, was over-expressed in the cluster A1. Cluster B was characterized among others by low expression levels of the gene *OTP*, which has been suggested previously as a prognostic marker for pulmonary carcinoids. This gene's expression also differed between the poor-prognosis and the good-prognosis atypical groups identified by the ML algorithm.

2.2.4 Conclusion and discussion

LNEN samples diagnosis is currently based on histopathological criteria. This study by integrating expression and methylation data identified molecular subgroups that were contrasted with the histopathological initial classification. On one hand, a supervised learning method, trained on these omics data to recognize the histopathological classification, was able to further divide the histopathological groups into molecular groups with different survival profiles. On the other hand, unsupervised analyses revealed molecular clusters which were further characterized and have potential clinical implications. Together these analyses allowed us to gain insights on the molecular characteristics of the lung neuroendocrine neoplasms, especially on the understudied atypical carcinoids samples.

Application of both supervised and unsupervised methods to omics data identified distinct lung neuroendocrine molecular profiles that do not exactly match the current histopathological classification, suggesting that molecular data could be beneficial for the diagnosis of these cancers. The supervised method, based on random forest, identified a subgroup of atypical carcinoids with a poor prognosis, similar to the prognosis of the aggressive LCNEC samples which could explain the so far observed intermediate survival of atypical carcinoids. Although these results revealed discrepancies between histopathological and molecular classification, they do not argue the relevance of pathological features that are still critical for cancer classification. Recent studies have shown, using computational histopathology, that histopathological features can be predictive of prognosis (Courtiol *et al.* 2019 [129]) as well as correlate with genomic alterations including structural variants and mutations in driver genes (Fu *et al.* 2019 [130]). Such studies indicate that histopathological and molecular data could complement each other, and further multi-disciplinary studies integrating those information would be beneficial for tumor diagnosis.

Using the unsupervised approach based on MOFA and clustering, different molecular clusters were identified. This analysis unveiled the supra-carcinoids exhibiting the molecular features of LCNECs while having the morphology of carcinoids. The

observation of such samples supports the previously proposed link between carcinoids and LCNEC samples [131, 94] and could inform future classifications of the lung neuroendocrine tumors. Also, the observation of intermediate cases could reflect potential transitional states between subtypes and if confirmed, inform on the progression of these diseases.

The unsupervised analysis also suggested the immune system and the retinoid and xenobiotic metabolism as biological pathways involved in pulmonary carcinoids. A molecular characterization of the different clusters identified as well potential targets that could influence the clinical management of the patient. However, one limitation of this study is that the molecular profile observed for one sample is not the molecular profile of the whole tumor but rather the profile of one piece of the tumor. The biomarker identified by such studies could not be representative of the whole tumor because of tumor heterogeneity. More and more spatial studies are being conducted and show that this heterogeneity should be considered for biomarker evaluation [132].

Another limitation of this study is related to the small samples size of the collected dataset, which is due to the rarity of the tumors studied. This limitation raises multiple issues. Firstly, it reduced the possibilities in model parameter tuning and thus potentially caused sub-optimal results. Also, the small sample size did not allow us to replicate our results. Further studies of larger sample sizes would be needed to confirm the existence of the new molecular clusters identified, especially the group of supra-carcinoids (around 5% in our series). One possibility to achieve this goal is to take advantage of all molecular studies already performed on the lung Neuroendocrine neoplasm (NEN) tumors by integrating the data and thus increasing the sample size of the datasets. However, integrating datasets resulting from different study designs can be challenging. In the following chapter 3, we address some of these challenges and propose a molecular map of the lung NEN samples that integrate molecular data from six different studies.

2.2.5 Contribution

In this chapter, my contribution focused on the supervised machine learning analysis, based on random forest and aiming at classifying the lung neuroendocrine tumors. This analysis identified molecular groups with different prognostic values. I contrasted the results from the supervised and unsupervised analyses. Finally, I took part in the discussions and interpretation of the other analyses conducted. Overall, I contributed to the generation of main figures, the redaction of the paper (mainly the sections related to the supervised analyses) and its reviewing process,

in particular with the addition of the supplementary Figures 9 to 12 and 27 available in the Annex A.

2.3 Article 1: Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-11276-9 OPEN

Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids

N. Alcala et al.#

The worldwide incidence of pulmonary carcinoids is increasing, but little is known about their molecular characteristics. Through machine learning and multi-omics factor analysis, we compare and contrast the genomic profiles of 116 pulmonary carcinoids (including 35 atypical), 75 large-cell neuroendocrine carcinomas (LCNEC), and 66 small-cell lung cancers. Here we report that the integrative analyses on 257 lung neuroendocrine neoplasms stratify atypical carcinoids into two prognostic groups with a 10-year overall survival of 88% and 27%, respectively. We identify therapeutically relevant molecular groups of pulmonary carcinoids, suggesting DLL3 and the immune system as candidate therapeutic targets; we confirm the value of *OTP* expression levels for the prognosis and diagnosis of these diseases, and we unveil the group of supra-carcinoids. This group comprises samples with carcinoid-like morphology yet the molecular and clinical features of the deadly LCNEC, further supporting the previously proposed molecular link between the low- and high-grade lung neuroendocrine neoplasms.

Correspondence and requests for materials should be addressed to L.F.-C. (email: fernandezcuestal@iarc.fr). #A full list of authors and their affiliations appears at the end of the paper.

According to the WHO classification from 2015[1] and a recent IARC-WHO expert consensus proposal[2], pulmonary carcinoids are low-grade typical and intermediate-grade atypical well-differentiated lung neuroendocrine tumours (LNETs) that belong to the group of lung neuroendocrine neoplasms (LNENs), which also includes the high-grade and poorly differentiated small-cell lung cancer (SCLC) and large-cell neuroendocrine carcinomas (LCNEC). Pulmonary carcinoids are rare malignant lesions, annual incidence of which has been increasing worldwide, especially at the advanced stages[3]. Pulmonary carcinoids account for 1–2% of all invasive lung malignancies: typical carcinoids exhibit good prognosis, although 10-23% metastasise to regional lymph nodes, resulting in a 5-year overall survival rate of 82–100%. The prognosis is worse for atypical carcinoids, with 40–50% presenting metastasis, reducing the 5-year overall survival rate to 50%.

Contrary to pulmonary carcinoids, most of which are eligible for upfront surgery at the time of diagnosis[3], LCNEC and SCLC require upfront aggressive, multimodal treatment for most of the patients. Owing to these differences in clinical management and prognosis, the accurate diagnosis of these diseases is critical. However, there is still no consensus on the optimal approach for their differential diagnosis;[2] the current criteria, based on morphological features and immunohistochemistry, are imperfect and inter-observer variations are common, especially when separating typical from atypical carcinoids[4], as well as atypical carcinoids from LCNEC in small biopsies[5]. Ki67 protein immune-reactivity has been suggested as a good marker of prognosis in LNENs as a whole, and for the differential diagnosis between carcinoids and SCLC[6,7], whereas this marker does not faithfully follow the defining histological criteria of typical and atypical carcinoids[4]. The difficulties in finding good markers to separate these diseases might be due to the limited amount of comprehensive genomic studies available for SCLC, LCNEC, and typical carcinoids, and the complete lack of such studies for atypical carcinoids[8]. In addition, such studies would also be needed to validate the recent proposed molecular link between pulmonary carcinoids and LCNEC[9,10].

In this study, we provide a comprehensive overview of the molecular traits of LNENs—with a particular focus on the understudied atypical carcinoids—in order to identify the mechanisms underlying the clinical differences between typical and atypical carcinoids, to understand the suggested molecular link between pulmonary carcinoids and LCNEC, and to find new candidates for the diagnosis and treatment of these diseases.

Results

Data. We have generated new data (genome, exome, transcriptome, and methylome) for 63 pulmonary carcinoids (including 27 atypical) and 20 LCNEC. In order to perform comparative analyses, we have reanalysed published data for 74 pulmonary carcinoids[11], 75 LCNEC[12], and 66 SCLC[13,14]. Taken together, we have performed multi-omics integrative analyses on 116 pulmonary carcinoids (including 35 atypical), 75 LCNEC, and 66 SCLC (Supplementary Fig. 1 and Supplementary Data 1).

Molecular groups of pulmonary carcinoids and LCNEC. We performed an unsupervised analysis of the expression and methylation data of the LNENs (i.e., 110 pulmonary carcinoids and 72 LCNEC) using the Multi-Omics Factor Analysis implementation of the group factor analysis statistical framework (Software MOFA)[15] (MOFA LNEN; Fig. 1a and Supplementary Figs. 2 and 3). We identified five latent factors explaining more than 2% of the variance in at least one data set, and among them, three latent factors provided consistent groups of samples with similar expression and methylation profiles (i.e., clusters). MOFA latent factors one (LF1) and two (LF2) explained a total of 45% and 34% of the variance in methylation and expression, respectively, and were both associated with survival (Supplementary Fig. 4). Using consensus clustering on these two latent factors (which explained most of the variation and thus carried most of the biological signal; Supplementary Figs. 5–7 and Supplementary Data 2–3), we identified three clusters, each of them enriched for samples of one of the three histopathological types (Fig. 1a). Cluster Carcinoid A was enriched for typical carcinoids (75%; Fisher's exact test p-value $< 2.2 \times 10^{-16}$); cluster Carcinoid B was enriched for atypical carcinoids (54%; Fisher's exact test p-value $< 2.2 \times 10^{-16}$) and male patients (79%; Fisher's exact test p-value $= 1.6 \times 10^{-9}$); and cluster LCNEC included 92% of the histopathological LCNEC (Fisher's exact test p-value $< 2.2 \times 10^{-16}$). Note that clustering based on LF1 to LF5, weighted by their proportion of variance explained, leads to the exact same clusters (Supplementary Fig. 8).

To assess whether the current histopathological classification could be improved by the combination of molecular and morphological characteristics, we undertook a machine-learning (ML) analysis. To do so, we combined the predictions from two independent random forest classifications, based on only-expression or only-methylation data. Using two independent models allowed the inclusion of samples for which only one of these data sets was available, thus maximising the power of subsequent analyses (Fig. 1b and Supplementary Fig. 9 for an alternative analysis based on both 'omic data sets simultaneously, but restricted to fewer samples). In order to avoid overfitting the data, we performed a leave-one-out cross-validation, with feature filtering and normalisation learned from the training set and applied to the test sample. To identify intermediate profiles, we defined a prediction category (unclassified) for samples that had a probability ratio between the two most probable classes close to one. We present in Fig. 1b the results for a cutoff ratio of 1.5, and show in Supplementary Fig. 10 the robustness of our results with regard to this ratio. Ninety-six per cent of the carcinoids predicted as typical by the ML were in cluster Carcinoid A (Fig. 1a). Similarly, the majority of ML-predicted atypical carcinoids (87%) belonged to cluster Carcinoid B.

We selected the ML-prediction groups with >10 samples (gathering the unclassified samples in one single group) and compared their overall survival using Cox's proportional hazard model (coloured groups in Fig. 1b). The machine learning trained on the histopathology stratified atypical carcinoids into two prognostic groups: the good-prognosis group (atypical reclassified as typical, in pink in Fig. 1b, c) with a 10-year overall survival similar to that of samples confirmed by ML as typical carcinoids (in black in Fig. 1b, c; 88% and 89%, respectively; Wald test p-value $= 0.650$); and the bad-prognosis group (atypical predicted as atypical, in red in Fig. 1b, c) with a 10-year overall survival similar to that of samples confirmed by ML as LCNEC (in blue in Fig. 1b, c; 27% and 19% respectively; Wald test p-value $= 0.574$; see also Supplementary Fig. 11). Machine-learning analyses based on other features -combined expression and methylation data (Supplementary Fig. 9), MOFA latent factors (Supplementary Fig. 12A), and Principal component analyses (PCA) principal components explaining more than 2% of the variance (Supplementary Fig. 12B)- led to qualitatively similar results.

Atypical carcinoids with LCNEC molecular characteristics. Six atypical carcinoids clustered with LCNEC in the MOFA LNEN (supra-carcinoids; Fig. 1a). Consistent with this clustering, this group displayed a survival similar to the other samples in the LCNEC cluster (10-year overall survival of 33% and 19%,

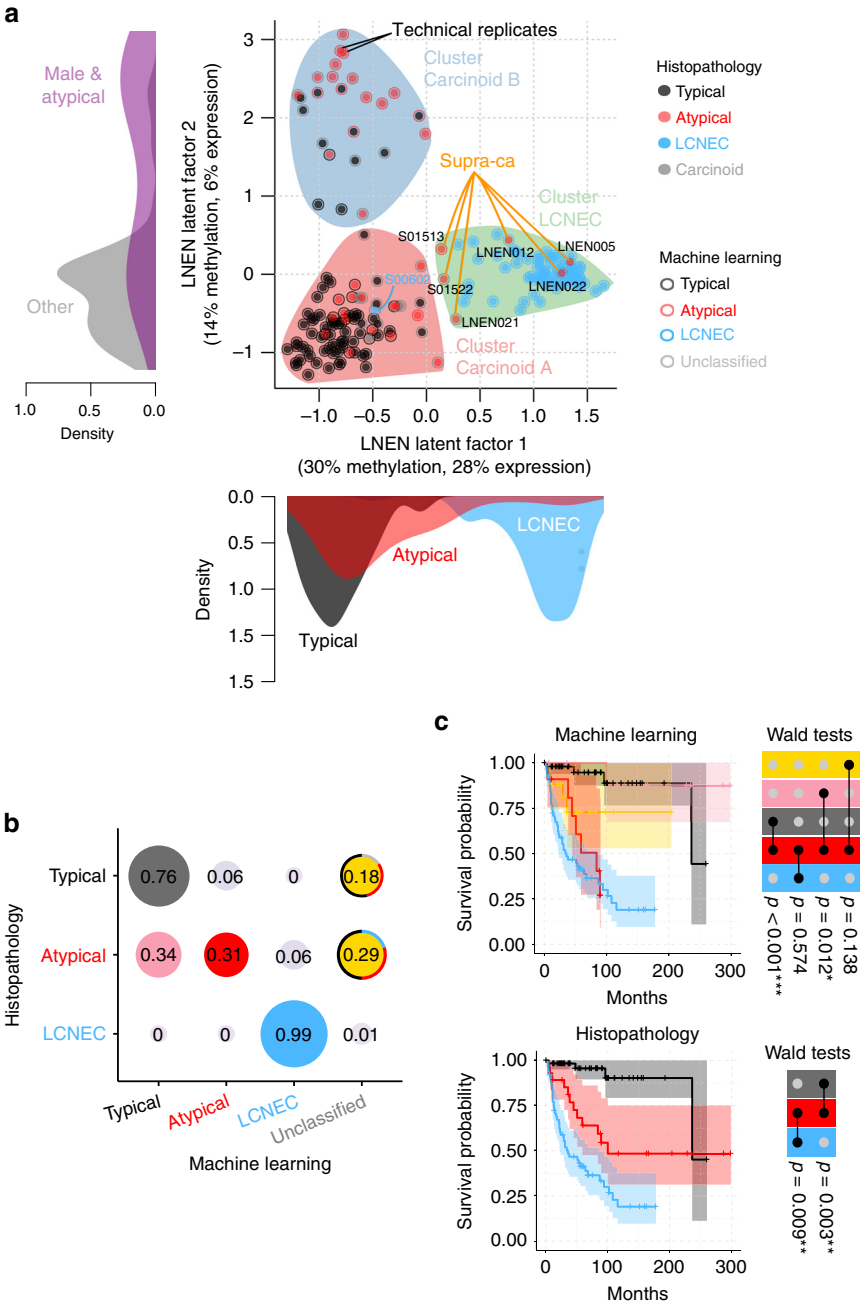


Fig. 1 Multi-omics (un)supervised analyses of lung neuroendocrine neoplasms. **a** Multi-omics factor analysis (MOFA) of transcriptomes and methylomes of LNEN samples (typical carcinoids, atypical carcinoids, and LCNEC). Point colours correspond to the histopathological types; coloured circles correspond to predictions of histopathological types by a machine learning (ML) algorithm (random forest classifier) outlined in **b**; filled coloured shapes represent the three molecular clusters identified by consensus clustering. The density of clinical variables that are significantly associated with a latent factor (ANOVA q-value < 0.05) are represented by kernel density plots next to each axis: histopathological type for latent factor 1, sex and histopathological type for latent factor 2. **b** Confusion matrix associated with the ML predictions represented on **a**. The different colours highlight the prediction groups considered in the survival analysis and the colours for machine learning are consistent between panel **b** and upper panel **c**. Black represents typical carcinoids predicted as typical, pink represents atypical carcinoids predicted as typical, red represents atypical carcinoids predicted as atypical, and blue represents LCNEC samples predicted as LCNEC. For the unclassified category, the most likely classes inferred from the ML algorithm are represented by coloured arcs (black for typical, red for atypical, blue for LCNEC, and light grey for discordant methylation-based and expression-based predictions). **c** Kaplan–Meier curves of overall survival of the different ML predictions groups (upper panel) and histopathological types (lower panel). Upper panel: colours of predicted groups match panel **b**. Lower panel: black-typical, red-atypical, blue-LCNEC. Next to each Kaplan–Meier plot, matrix layouts represent pairwise Wald tests between the reference group and the other groups, and the associated p-values; $0.01 \leq p < 0.05$, $0.001 \leq p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. Data necessary to reproduce the figure are provided in Supplementary Data 1

respectively; Wald test p-value = 0.574; Fig. 2a). The observed molecular link appears to be between supra-carcinoids and LCNEC rather than with SCLC, as shown by PCA and MOFA including expression data for 51 SCLC (Supplementary Figs. 6 and 13, respectively).

These samples originated from three different centres (two from each), and included two previously published samples (S01513 and S01522)[11], implying that this observation is unlikely to be the result of a batch effect. The limited number of supra-carcinoids did not allow to explore aetiological links; however, it is of note that one of them (LNEN005) belonged to a patient with professional exposure to asbestos (which is known to cause mesothelioma)[16] (Table 1), and the tumour harboured a splicing *BAP1* somatic mutation (a gene frequently altered in mesothelioma)[17]. This sample showed the highest mutational load (37 damaging somatic mutations; Supplementary Data 4). Gene set enrichment analyses (GSEA) of mutations in the hallmarks of cancer gene sets[18,19], showed a significant enrichment for the hallmark evading growth suppressor (q-value = 0.0213; Fig. 2b and Supplementary Data 5), while the hallmark genome instability and mutation was significant only at the 10% false discovery rate (FDR) threshold (q-value = 0.0970; Fig. 2b and Supplementary Data 5). We had access to the Haematoxylin and Eosin (H&E) stain for three of these supra-carcinoids, on which the pathologists discarded misclassifications with LCNEC, SCLC, or mesothelioma in the case of the asbestos-exposed *BAP1*-mutated sample (Fig. 2c and Table 1).

While generally similar to LCNEC, and albeit based on small numbers, the supra-carcinoids appeared to have nonetheless some distinct genomic features based on genome-wide expression and methylation profiles (Fig. 2d). Supra-carcinoids displayed higher levels of immune checkpoint genes (both receptors and ligands; Fig. 2e), and also harboured generally higher expression levels of MHC class I and II genes (Fig. 2e and Supplementary Fig. 14). Interestingly, the interferon-gamma gene—a prominent immune-stimulator, in particular of the MHC class I and II genes—also showed high-expression levels in these samples (Supplementary Fig. 14). The differences in immune checkpoint gene expression levels between groups were not explained by the amount of infiltrating cells, as estimated by deconvolution of gene expression data with software quanTIseq (Fig. 2f, left panel). However, supra-carcinoids contained the highest levels of neutrophils (greater than the 3rd quartile of the distributions of neutrophils in the other groups; Fig. 2f, right panel). Permutation tests showed that these levels were significantly higher than in other carcinoid groups and in SCLC, but not than in LCNEC (Supplementary Fig. 15). Concordantly, GSEA showed that MOFA LNEN LF1 (separating LCNEC and supra-carcinoids from the other carcinoids) was significantly associated with neutrophil chemotaxis and degranulation pathways (Supplementary Data 6). By contrast, no such association was observed in the MOFA performed only on carcinoids and SCLC samples (Supplementary Figs. 6C and 13C and Supplementary Data 6).

Mutational patterns of pulmonary carcinoids. In a previous study, mainly including typical carcinoids, we detected *MEN1*, *ARID1A*, and *EIF1AX* as significantly mutated genes[11]. We also found that covalent histone modifiers and subunits of the SWI/SNF complex were mutated in 40% and 22.2% of the cases, respectively. Genomic alterations in these genes and pathways were also seen in the new samples included in this study (Fig. 3a, Supplementary Fig. 16, and Supplementary Data 4). Apart from the above-mentioned genes, *ATM*, *PSIP1*, and *ROBO1* also showed some evidence, among others, for recurrent mutations in pulmonary carcinoids (Fig. 3a). In addition to point mutations and small indels, the *ARID2*, *DOT1L*, and *ROBO1* genes were also altered by chimeric transcripts (Fig. 3b). *MEN1* was also inactivated by genomic rearrangement in a carcinoid sample with a chromothripsis pattern affecting chromosomes 11 and 20 (Fig. 3c). The full lists of somatically altered genes, chimeric transcripts, and genomic rearrangements are presented in Supplementary Data 4, 7, and 8, respectively. Of note, *MEN1* mutations were significantly associated with the atypical carcinoid histopathological subtype (Fisher's exact test p-value = 0.0096), as well as MOFA LNEN LF2.

Altered pathways in pulmonary carcinoids. The third latent factor from the MOFA LNEN accounted for 8% and 6% of the variance in expression and methylation, respectively, but unlike LF1 and LF2, LF3 was not associated with patient survival (Supplementary Fig. 4). The molecular variation explained by LF3 appeared to capture different molecular profiles within cluster Carcinoid A (Supplementary Fig. 13B). We therefore undertook an additional MOFA restricted to pulmonary carcinoid samples only (MOFA LNET; Fig. 4a and Supplementary Fig. 17). This MOFA identified five latent factors that explained at least 2% of the variance in one data set. As expected, the first two latent factors of the MOFA LNET were highly correlated with LF2 and LF3 from the MOFA LNEN, respectively, (Pearson correlation >0.96; Supplementary Fig. 13B), and explained 41% and 35% of the variance in methylation and expression, respectively. Integrative consensus clustering using LF1 and LF2 of the MOFA LNET identified three clusters (Supplementary Fig. 18): cluster Carcinoid A1 and cluster Carcinoid A2, that together correspond to the samples in cluster Carcinoid A of the MOFA LNEN, plus the supra-carcinoids; and cluster Carcinoid B (as for the clustering of LNEN samples, a clustering based on LF1-LF5 weighted by their proportion of variance explained, led to the exact same clusters; Supplementary Fig. 8). LF2 was associated with age, with cluster Carcinoid A1 enriched for older patients ((60, 90] years old) and cluster Carcinoid A2 enriched for younger patients ((15, 60] years old).

We applied GSEA to identify the pathways associated with the different latent factors. We found significant associations with the immune system and the retinoid and xenobiotic metabolism pathways (Supplementary Data 6). Numerous Gene Ontology (GO) terms and KEGG pathways were related to the immune system, immune cell migration, and infectious diseases. The GO terms and KEGG pathways related to immune cell migration included leucocyte migration, chemotaxis, cytokines, and interleukin 17 signalling. In particular, the expression of all β-chemokines (including CCL2, CCL7, CCL19, CCL21, CCL22, known to attract monocytes and dendritic cells)[20] (Supplementary Data 6), and all CXC chemokines (such as IL8, CXCL1, CXCL3, and CXCL5, known to attract neutrophils)[21], were positively correlated with MOFA LNEN LF1 (separating pulmonary carcinoids from LCNEC) and negatively correlated with MOFA LNET LF2 (separating clusters Carcinoid A1 and A2).

The different LNET clusters did not differ in their total amounts of estimated proportions of immune cells, but they did differ in their composition (Supplementary Fig. 19): cluster Carcinoid A (particularly A1) was significantly enriched in dendritic cells, and cluster Carcinoid B, in monocytes (Fig. 4b, upper panel). As monocytes can differentiate into dendritic cells in a favourable environment[22], we assessed the levels of *LAMP3* and *CD1A* dendritic-cells markers[23], and found that samples in cluster Carcinoid A1 presented high-expression levels of these genes (Fig. 4b, lower panel), implying that this cluster was indeed enriched for dendritic cells. We pursued this further by assessing

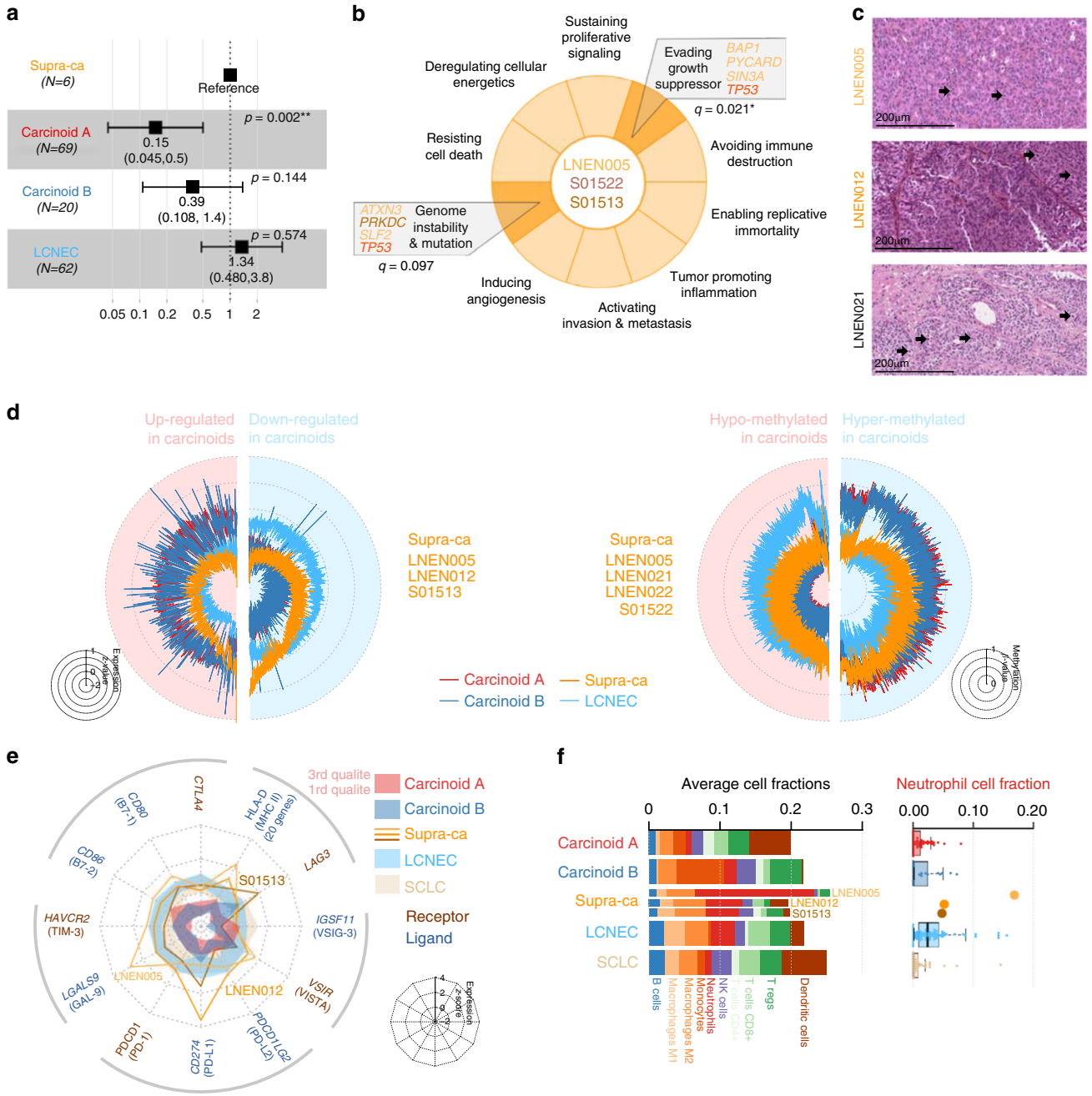


Fig. 2 Molecular characterisation of supra-carcinoids. **a** Forest plot of hazard ratios for overall survival of the supra-carcinoids, compared to Carcinoid A and B, and LCNEC. The number of samples (*N*) in each group is given in brackets. The black box represent estimated hazard ratios and whiskers represent the associated 95% confidence intervals. Wald test *p*-values are shown on the right. **b** Enrichment of hallmarks of cancer for somatic mutations in supra-carcinoids. Dark colours highlight significantly enriched hallmarks at the 10% false discovery rate threshold; corresponding mutated genes are listed in the boxes, and enrichment *q*-values are reported below. **c** Hematoxylin and Eosin (H&E) stains of three supra-carcinoids. In all cases, an organoid architecture with tumour cells arranged in lobules or nests, forming perivascular palisades and rosettes is observed; original magnification x200. Arrows indicate mitoses. **d** Radar charts of expression and methylation levels. Each radius corresponds to a feature (gene or CpG site), with low values close to the centre and high values close to the edge. Coloured lines represent the mean of each group. Left panel: expression *z*-scores of genes differentially expressed between clusters Carcinoid A and LCNEC or between Carcinoid B and LCNEC. Right panel: methylation *β*-values of differentially methylated positions between Carcinoid A and LCNEC clusters or between Carcinoid B and LCNEC clusters. **e** Radar chart of the expression *z*-scores of immune checkpoint genes (ligands and receptors) of each group. **f** Left panel: average proportion of immune cells in the tumour sample for each group, as estimated from transcriptomic data using software quanTIseq. Right panel: boxplot and beeswarm plot (coloured points) of the estimated proportion of neutrophils, where centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. Data necessary to reproduce the figure are provided in Supplementary Data 1, 4, 5, 12, 17, and in the European Genome-phenome Archive

Table 1 Characteristics of supra-carcinoids

	LNEN005	LNEN012	LNEN021	LNEN022	S01513	S01522
Classification						
Histopathology	Atypical	Atypical	Atypical	Atypical	Atypical	Atypical
Morphological characteristics	Carcinoid morph. 2 mitoses/2 mm^2 No necrosis	Carcinoid morph. 2 mitoses/2 mm^2 No necrosis	LCNEC morph. 4 mitoses/2 mm^2 No necrosis	NA	NA	NA
Machine learning	LCNEC	LCNEC	Unclassified	Unclassified	Atypical	Unclassified
Clinical data						
Sex	Male	Female	Female	Female	Male	Male
Age at diagnosis	80	70	83	58	58	63
TNM Stage	IB	IIIC	IA1	IIB	IIIA	IV
Overall survival (months)	144.6	111.7	29.8	36.1	59	7
Epidemiology						
Smoking status	Former	NA	NA	NA	Never	Current
Other known exposure	Asbestos	NA	NA	NA	NA	NA
Multi-omics data						
Data available	WES, RNAseq, Epic 850K	RNAseq	Epic 850K	Epic 850K	WGS, RNAseq	WES, Epic 850K
Cluster MOFA LNEN	LCNEC	LCNEC	LCNEC	LCNEC	LCNEC	LCNEC
Cluster MOFA LNET	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1	Carcinoid A1
Selected mutated genes	*JMJD1C, KDM5C, BAP1*	NA	NA	NA	*DNAH17*	*TP53*
Mean FPKM of IC genes[a]	8.12	10.32	NA	NA	3.15	NA
MKI67 FPKM	2.6	7.3	NA	NA	1.9	NA

FPKM refers to Fragments Per Kilobase per Million reads. The median FPKM of immune checkpoint (IC) genes was calculated based on the genes included in Fig. 2e, excluding HLA genes because of their very large expression levels
[a]IC genes median FPKM values for pulmonary carcinoids, LCNEC and SCLC are 1.0, 3.5, and 3.2, respectively

the CD1A protein levels by immunohistochemistry (IHC) in an independent series of pulmonary carcinoids, and found that 60% of them (12 out of 20) were enriched in CDA1-positive dendritic cells, confirming the presence of dendritic cells in a subgroup of pulmonary carcinoids (Fig. 4c and Supplementary Data 9).

Regarding the retinoid and xenobiotic metabolism pathways (e.g., elimination of drugs and environmental pollutants), the main genes driving the correlation with MOFA latent factors were the phase II enzymes involved in glucuronosyl-transferase activity (Supplementary Data 6), but also the phase I cytochrome P450 (CYP) proteins. These pathways were positively correlated with MOFA LNEN LF2 (separating LNEN clusters A and B) and negatively correlated with MOFA LNET LF1 (separating LNET clusters A1 and A2 from cluster B). Indeed, we found that samples in cluster Carcinoid B were characterised by high levels of the CYP family of genes, and a very strong expression of several UDP glucuronosyl-transferases *UGT* genes (median FPKM = 4.6 in *UGT2A3* and 28.1 in *UGT2B* genes; Fig. 4d), which contrasts with the low levels in other carcinoids (median FPKM = 0 for both *UGT2A3* and *UGT2B*; Fig. 4d), LCNEC (median FPKM = 0 and 1.2 for *UGT2A3* and *UGT2B*; Supplementary Fig. 20) and SCLC (median FPKM = 0 and 0.3 for *UGT2A3* and *UGT2B*; Supplementary Fig. 20).

Molecular groups of pulmonary carcinoids. We explored the molecular characteristics of each cluster from the MOFA LNET based on their core differentially expressed coding genes (core-DEGs, the expression levels of which defined a given group of samples), corresponding promoter methylation profiles (Fig. 5a and Supplementary Data 10), and their somatic mutational patterns (Figs. 3a and 4a). To achieve this goal, we computed the DEGs in all pairwise comparisons between a focal group and the other groups, and then defined core-DEGs as the intersection of the resulting gene sets. We show in Supplementary Fig. 21 that core-DEGs are almost exclusively a subset of the DEGs between the focal group and samples from all other groups taken together. We correlated the gene expression and promoter methylation data of the core-DEGs to identify genes, which expression could be mainly explained by their methylation patterns (Fig. 5a). One of the top correlations was found for *HNF1A* and *HNF4A* homeobox genes (Supplementary Fig. 22), which were strongly downregulated in cluster Carcinoid A1 samples (Supplementary Fig. 23). In addition, the promoter regions of these genes also harboured core-DMPs (differentially methylated positions) of cluster Carcinoid A1, indicating that their methylation profile is specific of this cluster (Supplementary Data 11). These two genes have been reported as having a role in the transcriptional regulation of *ANGPTL3*, CYP, and UGT genes[24], and could thus explain the differential expression of these genes between the clusters. Samples in cluster Carcinoid A1 were also characterised by high-expression levels of the delta like canonical Notch ligand 3 (*DLL3*, 75% with FPKM > 1) and its activator the achaete-scute family bHLH transcription factor 1 (*ASCL1*) (Fig. 5a and Supplementary Data 10), similar to SCLC and LCNEC (Fig. 5b); however, the expression levels of NOTCH genes did not differ between the different groups (Supplementary Fig. 24). The supra-carcinoids were negative for *DLL3* expression (Fig. 5b), and had generally high-expression levels of *NOTCH1-3* (Supplementary Fig. 24). We additionally tested the DLL3 protein levels in the aforementioned independent series of 20 pulmonary carcinoids and found 40% (eight out of 20) with relatively high expression of DLL3 (Fig. 4d and Supplementary Data 9), while in the other 12 samples DLL3 was strikingly absent (Fig. 4d and Supplementary Data 9). Furthermore, we found a correlation between the protein levels of DLL3 and CD1A (Pearson test p-value = 0.00034; Supplementary Fig. 25), providing additional evidence for the existence of a DLL3+ CD1A+ subgroup of carcinoids. Core-DEGs in cluster Carcinoid A2 included the low levels of *SLIT1* (slit guidance ligand 1; 97% with FPKM < 0.01), and *ROBO1* (roundabout guidance receptor 1; 56% with FPKM < 1) (Fig. 5a, b and Supplementary Data 10). This cluster also contained the four samples with somatic mutations in the eukaryotic translation initiation factor 1A X-linked (*EIF1AX*) gene (Fig. 4a). Concordantly, samples with *EIF1AX* mutations had significantly higher coordinates on the MOFA LNET LF2 (t-test p-value = 0.0342).

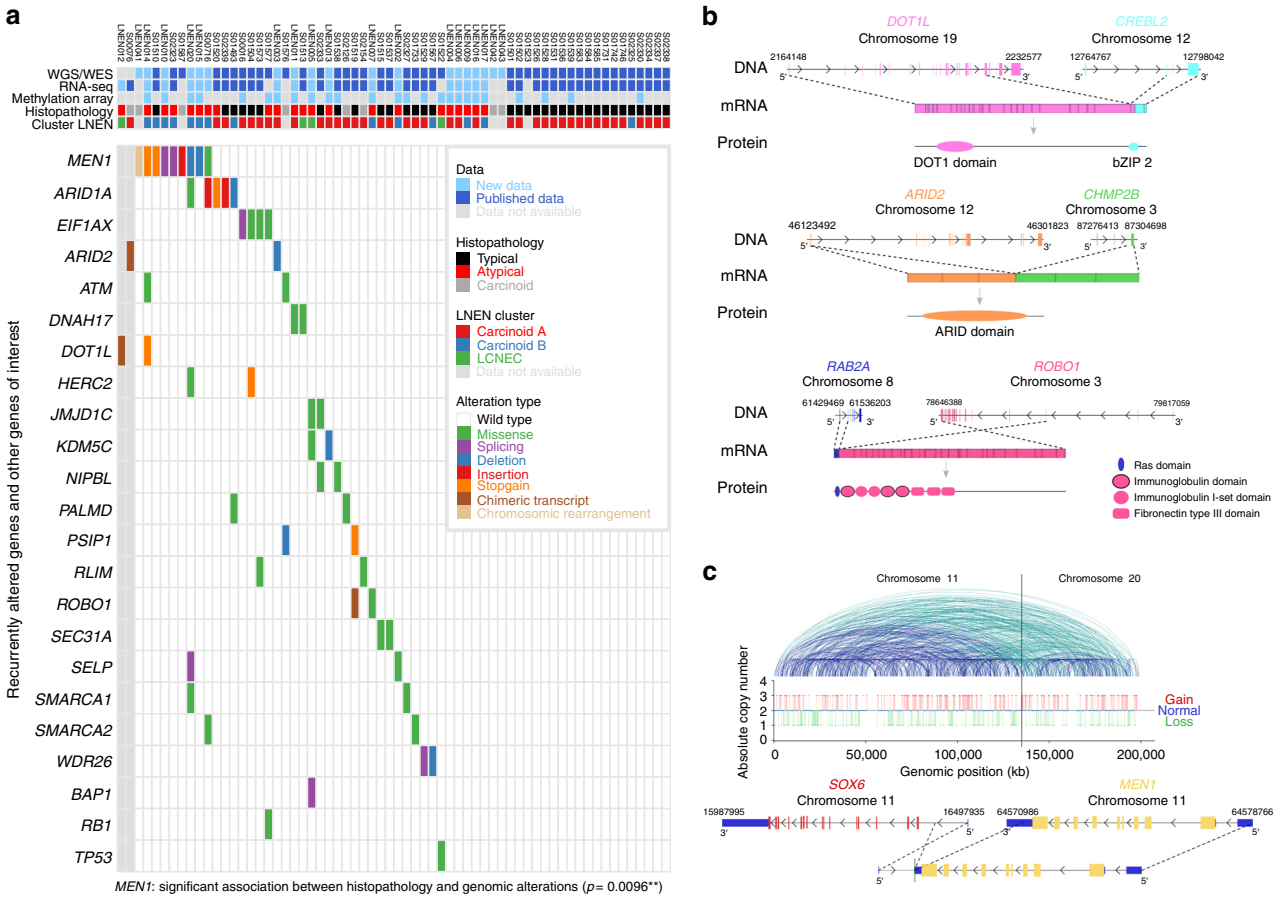


Fig. 3 Mutational patterns of pulmonary carcinoids. **a** Recurrent and cancer-relevant altered genes found in pulmonary carcinoids by WGS and WES. Fisher's exact test *p*-value for the association between *MEN1* and the atypical carcinoid histopathological subtype is given in brackets; $0.01 \le p < 0.05$, $0.001 \le p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. **b** Chimeric transcripts affecting the protein product of *DOT1L* (upper panel), *ARID2* (middle panel), and *ROBO1* (lower panel). For each chimeric transcript the DNA row represents genes with their genomic coordinates, the mRNA row represents the chimeric transcript, and the protein row represents the predicted fusion protein. **c** Chromotripsis case LNEN041, including an inter-chromosomic rearrangement between genes *MEN1* and *SOX6*. Upper panel: copy number as a function of the genomic coordinates on chromosomes 11 and 20; a solid line separates chromosomes 11 and 20. Blue and green lines depict intra- and inter-chromosomic rearrangements, respectively. Lower panel: *MEN1* chromosomic rearrangement observed in this chromotripsis case. Data necessary to reproduce the figure are provided in Supplementary Data 4, 7, and 8

As expected based on Fig. 4d, several UGT genes were core-DEGs of cluster Carcinoid B (Fig. 5a). Also, accordingly with the worse survival of patients in this cluster (Fig. 2a), these samples were also characterised by the expression of angiopoietin like 3 (*ANGPTL3*, 90% with FPKM > 1), and the erb-b2 receptor tyrosine kinase 4 (*ERBB4*, 67% with FPKM > 1) (Fig. 5b). This cluster was also characterised by the universal downregulation of orthopedia homeobox (*OTP*; 90% with FPKM < 1), and NK2 homeobox 1 (*NKX2-1*; 90% FPKM < 1) (Fig. 5b). Interestingly, the SCLC-combined LCNEC sample (S00602) that clustered with the pulmonary carcinoids in the MOFA LNEN (Fig. 1a) was the only LCNEC in our series harbouring high-expression levels of *OTP* (290.26 FPKM vs. 9.89 FPKM for the 2nd highest within LCNEC, the median for LCNEC being 0.22 FPKM). *UGT* genes, *ANGPTL3*, and *ERBB4* were also core-DEGs of cluster B samples when compared to LNEN clusters Carcinoid A and LCNEC (Supplementary Data 12), which indicates that their expression levels also significantly differed from that of LCNEC. Cluster Carcinoid B included all observed *MEN1* mutations, which is consistent with the fact that samples with *MEN1* mutations had significantly lower coordinates on the MOFA LNET LF1 (*t*-test *p*-value $= 7 \times 10^{-6}$; Fig. 4a). Nevertheless, mutations in this gene did not explain the poorer prognosis of this group of samples compared to other LNET (logrank *p*-value > 0.05; Supplementary Fig. 26). To gain some insights into what might be driving the bad prognosis of cluster Carcinoid B samples, we performed a GSEA of mutations in hallmarks of cancer gene sets[18,19]; while clusters Carcinoid A1 and A2 were not enriched for any hallmark of cancer, cluster Carcinoid B was significantly enriched for genes involved in evading growth suppressor, sustaining proliferative signalling, and genome instability and mutation at the 5% FDR (Fig. 5c). We also performed a Cox regression with elastic net regularisation based on the core-DEGs of this cluster; the model selected eight coding genes explaining the overall survival, *OTP* being one of them (Fig. 5d and Supplementary Data 13). Further supporting their prognostic value, we found that the expression of four of these genes was significantly different between the good- and the poor-prognosis atypical carcinoids based on the machine-learning predictions (Fig. 1c, upper panel and Supplementary Fig. 27).

Finally, we also checked the *MKI67* expression levels in the different molecular groups and found relatively low levels in the clusters Carcinoids A1, A2, and B (78% with FPKM < 1) and high levels in the supra-carcinoids (FPKM > 1 in the three samples). As

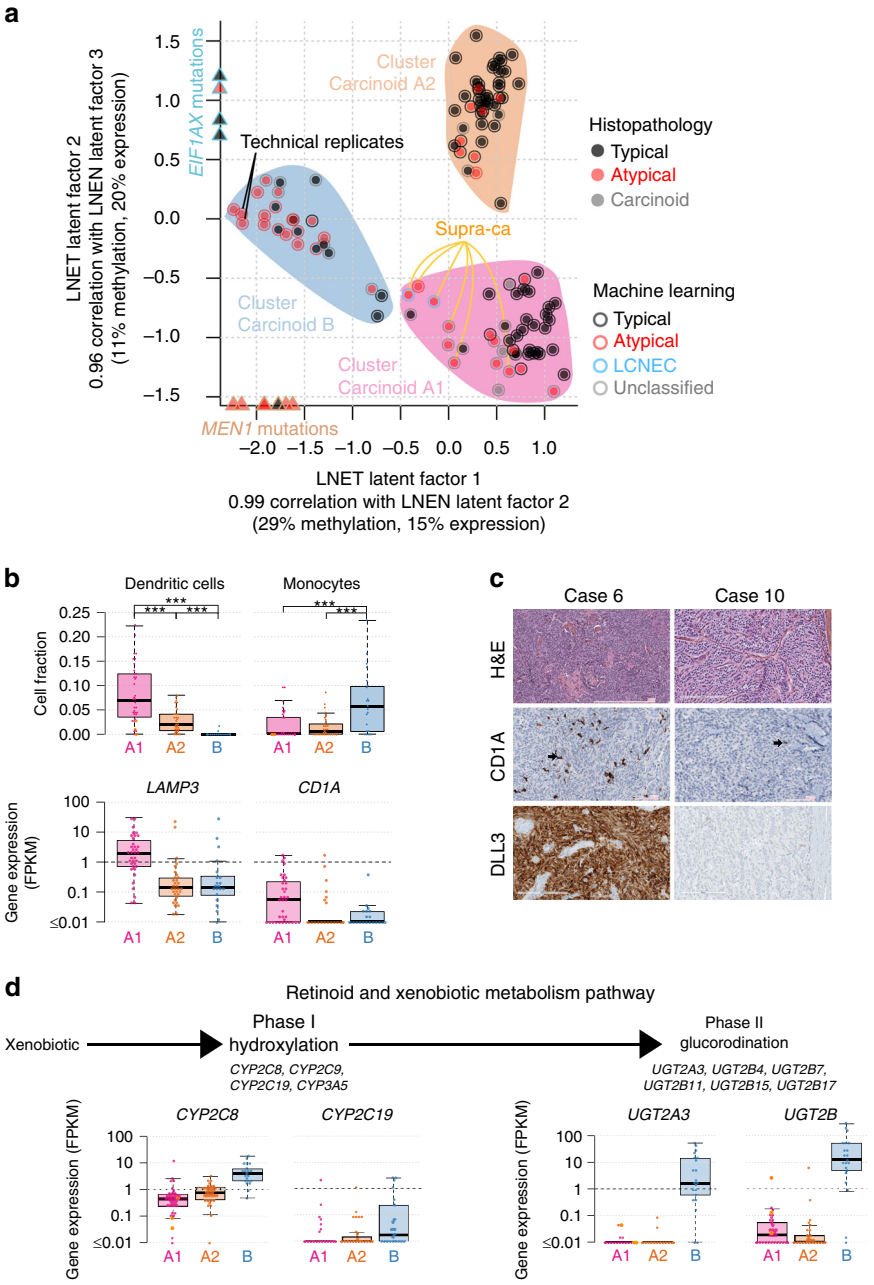


Fig. 4 Multi-omics unsupervised analysis of lung neuroendocrine tumours. **a** Multi-omics factor analysis (MOFA) of transcriptomes and methylomes restricted to LNET samples (pulmonary carcinoids). Design follows that of Fig. 1a; filled coloured shapes represent the three molecular clusters (Carcinoid A1, A2, and B) identified by consensus clustering. The position of samples harbouring mutations significantly associated with a latent factor (ANOVA q-value < 0.05) are highlighted by coloured triangles on the axes. **b** Upper panel: boxplots of the proportion of dendritic cells in the different molecular clusters (Carcinoid A1, A2, and B) and the supra-carcinoids, estimated from transcriptomic data using quanTIseq (Methods). The permutation test q-value range is given above each comparison: q-value < 0.001 is annotated by three stars. Lower panel: boxplots of the expression levels of *LAMP3* (CDLAMP) and *CD1A*. **c** DLL3 and CD1A immunohistochemistry of two typical carcinoids: case 6 (DLL3+ and CD1A+), and case 10 (DLL3- and CD1A-). Upper panels: Hematoxylin & Eosin Saffron (H&E) stain. Middle panels: staining with CD1 rabbit monoclonal antibody (cl EP3622; VENTANA), where arrows show positive stainings. Lower panels: Staining with DLL3 assay (SP347; VENTANA). **d** Expression levels of genes from the retinoid and xenobiotic metabolism pathway—the most significantly associated with MOFA latent factor 1—in the different molecular clusters. Upper panel: schematic representation of the phases of the pathway. Lower panel: boxplot of expression levels of *CYP2C8* and *CYP2C19* (both from the CYP2C gene cluster on chromosome 10), *UGT2A3,* and the total expression of *UGT2B* genes (from the UGT2 gene cluster on chromosome 4), expressed in fragments per kilobase million (FPKM) units. In all panels, boxplot centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. Data necessary to reproduce the figure are provided in Supplementary Data 1, 4, 9, and in the European Genome-phenome Archive

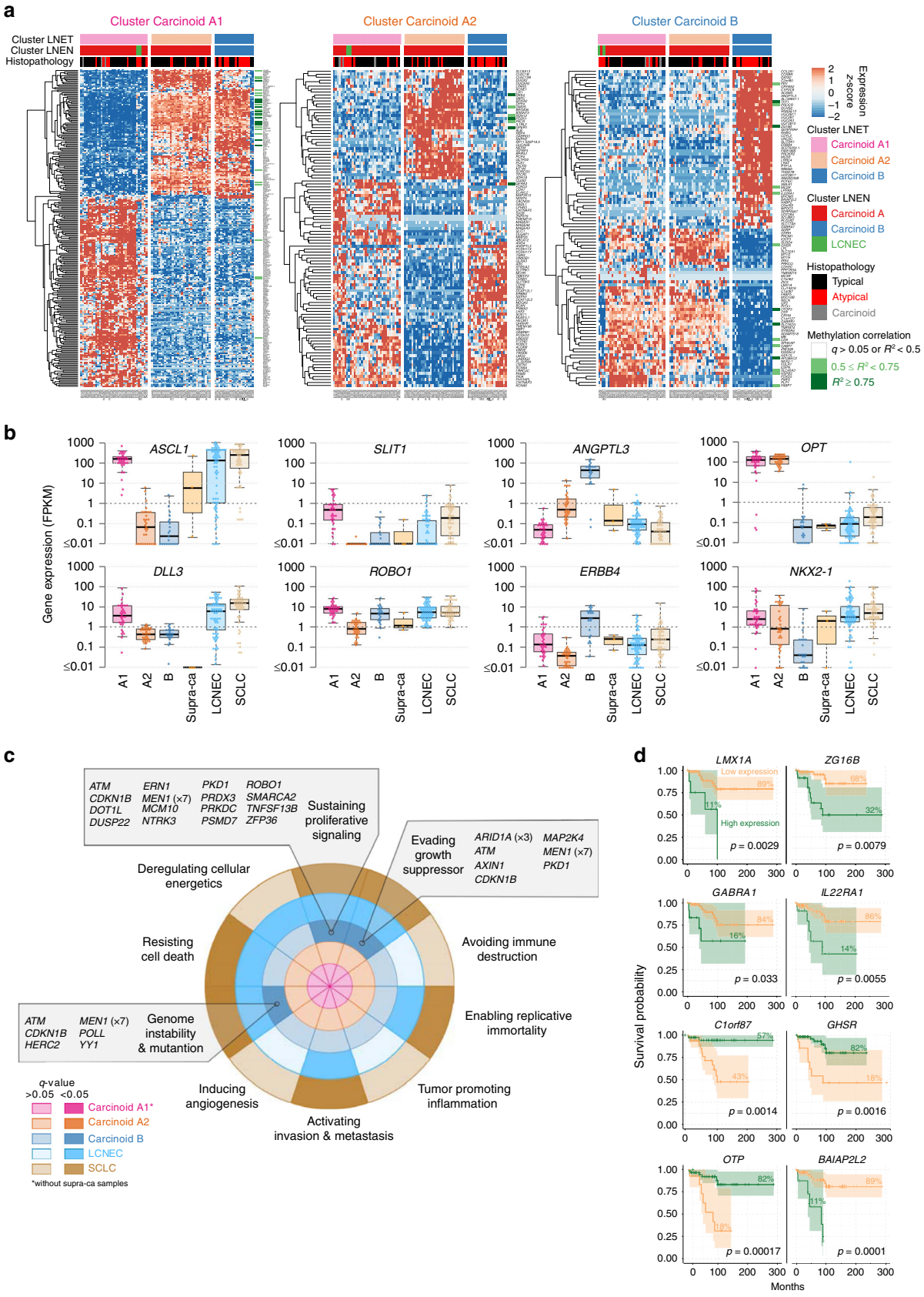

a
Cluster Carcinoid A1
Cluster Carcinoid A2
Cluster Carcinoid B
Cluster LNET
Cluster LNEN
Histopathology
Expression z-score
Cluster LNET
Carcinoid A1
Carcinoid A2
Carcinoid B
Cluster LNEN
Carcinoid A
Carcinoid B
LCNEC
Histopathology
Typical
Atypical
Carcinoid
Methylation correlation
q > 0.05 or R² < 0.5
0.5 ≤ R² < 0.75
R² ≥ 0.75
b
Gene expression (FPKM)
ASCL1
SLIT1
ANGPTL3
OPT
DLL3
ROBO1
ERBB4
NKX2-1
A1
A2
B
Supra-ca
LCNEC
SCLC
c
ATM CDKN1B DOT1L DUSP22
ERN1 MEN1 (×7) MCM10 NTRK3
PKD1 PRDX3 PRKDC PSMD7
ROBO1 SMARCA2 TNFSF13B ZFP36
Sustaining proliferative signaling
Evading growth suppressor
ARID1A (×3) ATM AXIN1 CDKN1B
MAP2K4 MEN1 (×7) PKD1
Avoiding immune destruction
Enabling replicative immortality
Tumor promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutantion
ATM CDKN1B HERC2
MEN1 (×7) POLL YY1
Resisting cell death
Deregulating cellular energetics
q-value
>0.05 <0.05
Carcinoid A1*
Carcinoid A2
Carcinoid B
LCNEC
SCLC
*without supra-ca samples
d
LMX1A
ZG16B
GABRA1
IL22RA1
C1orf87
GHSR
OTP
BAIAP2L2
Low expression
High expression
p = 0.0029
p = 0.0079
p = 0.033
p = 0.0055
p = 0.0014
p = 0.0016
p = 0.00017
p = 0.0001
Survival probability
Months

Fig. 5 Molecular groups of pulmonary carcinoids. **a** Heatmaps of the expression of core differentially expressed genes of each molecular cluster, i.e., genes that are differentially expressed in all pairwise comparisons between a focal cluster and the other clusters. Green bars at the right of each heatmap indicate a significant negative correlation with the methylation level of at least one CpG site from the gene promoter region. The colour scale depends on the range of *q*-value (*q*) and squared correlation estimate (R^2) of the correlation test. **b** Boxplots of the expression levels of selected cancer-relevant core genes, in fragment per kilobase million (FPKM) units, where centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. **c** Characteristic hallmarks of cancer in each molecular cluster (Carcinoid A1 without the supra-carcinoids, A2, and B), LCNEC, and SCLC. Coloured concentric circles correspond to the molecular clusters. For each cluster, dark colours highlight significantly enriched hallmarks (Fisher's exact test *q*-value < 0.05). The mutated genes contributing to a given hallmark are listed in the boxes. Recurrently mutated genes are indicated in brackets by the number of samples harbouring a mutation. **d** Survival analysis of pulmonary carcinoids based on the expression level of eight core genes of cluster Carcinoid B. The genes were selected using a regularised GLM on expression data. For each gene, coloured lines correspond to the Kaplan–Meier curve of overall survival for individuals with a high (green) and low (orange) expression level of this gene. Cutoffs for the two groups were determined using maximally selected rank statistics (Methods). The percentage of samples in each group is represented above each Kaplan–Meier curve and the logrank test *p*-value is given in bottom right for each gene. Data necessary to reproduce the figure are provided in Supplementary Data 5, 10, and in the European Genome-phenome Archive

expected, LCNECs and SCLCs carried high levels of this gene (FPKM > 1 in 99% and 92% of the samples, respectively). Although the levels of *MKI67* for each of the clusters were different, further analyses showed that *MKI67* expression levels alone were not able to accurately separate good- from poor-prognosis pulmonary carcinoids (Supplementary Fig. 11B, C).

An overview of the different molecular groups of pulmonary carcinoids and their most relevant characteristics is displayed in Fig. 6.

Discussion

Lung neuroendocrine neoplasms are a heterogeneous group of tumours with variable clinical outcomes. Here, we characterised and contrasted their molecular profiles through integrative analysis of transcriptome and methylome data, using both machine-learning (ML) techniques and multi-omics factor analyses (MOFA). ML analyses showed that the molecular profiles could distinguish survival outcomes within patients with atypical carcinoid morphological features, splitting them into patients with good typical-carcinoid-like survival and patients with a clinical outcome similar to LCNEC. Overall, out of the 35 histopathologically atypical carcinoids, ML reclassified 12 into the typical category.

Unsupervised MOFA and subsequent gene-set enrichment analyses unveiled the immune system and the retinoid and xenobiotic metabolism as key deregulated processes in pulmonary carcinoids, and identified three molecular groups—clusters—with clinical implications (Fig. 6). The first group (cluster A1) presented high infiltration by dendritic cells, which are believed to promote the recruitment of immune effector cells resulting in a strongly active immunity[25]. Samples in cluster A1 showed overexpression of *ASCL1* and *DLL3*. The transcription factor ASCL1 is a master regulator that induces neuronal and neuroendocrine differentiation. It regulates the expression of *DLL3*, which encodes an inhibitor of the Notch pathway[26]. Overexpression of *ASCL1* and *DLL3* is a characteristic of the SCLC of the classic subtype[26] and of type-I LCNEC[12]. We validated the expression of DLL3 in an independent series of 20 pulmonary carcinoids assessed by immunohistochemistry (IHC; 40% positive). The fact that we found a correlation between the protein levels of DLL3 and CD1A (a marker of dendritic cells also assessed by IHC in this series; 60% positive) provides orthogonal evidence to support the existence of this molecular group. Phase I trials have provided evidence for clinical activity of the anti-DLL3 humanised monoclonal antibody in high–DLL3-expressing SCLCs and LCNECs[27], and additional clinical trials are ongoing in other cancer types.

The second group (cluster A2) harboured recurrent somatic mutations in *EIF1AX*, and showed downregulation of the *SLIT1* and *ROBO1* genes. SLIT and ROBO proteins are known to be axon-guidance molecules involved in the development of the nervous system[28], but the SLIT/ROBO signalling has also been associated with cancer development, progression, and metastasis. Pulmonary neuroendocrine cells (PNEC) represent 1% of the total lung epithelial cell population[29], they reside isolated (Kultchinsky cells) or in clusters named neuroepithelial bodies (NEBs), and are believed to be the cell of origin of most lung neuroendocrine neoplasms[30]. In the normal lung, it has been shown that *ROBO1/2* are expressed, exclusively, in the PNECs, and that the SLIT/ROBO signalling is required for PNEC assembly and maintenance in NEBs[31]. In cancer, this pathway mainly suppresses tumour progression by regulating invasion, migration, and apoptosis, and therefore, is often downregulated in many cancer types[28]. More specifically, the SLIT1/ROBO1 interaction can inhibit cell invasion by inhibiting the SDF1/CXCR4 axis, and can attenuate cell cycle progression by destruction of β-catenin and CDC42[28]. Potential clinical avenues to this finding exist, especially the ongoing development of CXCR4 inhibitors.

The third molecular group (cluster B) was enriched in monocytes and depleted of dendritic cells, and had the worst median survival. Even in the presence of T cell infiltration, this immune contexture suggests an inactive immune response, dominated by monocytes and macrophages with potent immunosuppressive functions, and almost devoid of the most potent antigen-presenting cells, dendritic cells, suggesting dendritic cell-based immunotherapy as a therapeutic option for this group of samples[32]. Cluster B was also characterised by recurrent somatic mutations in *MEN1*, the most frequently altered gene in pulmonary carcinoids and pancreatic NETs[33], which is in line with the common embryologic origin of pancreas and lung. *MEN1* was inactivated by genomic rearrangement due to a chromothripsis event affecting chromosomes 11 and 20 in one of our samples. This observation, together with two additional reported cases involving chromosomes 2, 12, and 13[11], and chromosomes 2, 11, and 20[34], respectively, suggest that chromothripsis is a rare but recurrent event in pulmonary carcinoids. Interestingly, *MEN1* mutations did not have a clear prognostic value in our series. Regarding the above-mentioned deregulation of the retinoid and xenobiotic metabolism in pulmonary carcinoids, samples in cluster B presented high levels of UGT and CYP genes. In line with previous studies[35,36], these samples also harboured low levels of *OTP*, which gene expression levels were correlated with survival in the ML predictions. High levels of *ANGPTL3* and *ERBB4* were also detected in this group of samples, representing candidate therapeutic opportunities. ANGPTL3 is involved in new blood vessel growth and stimulation of the MAPK pathway[37]. This protein has been found aberrantly expressed in several types

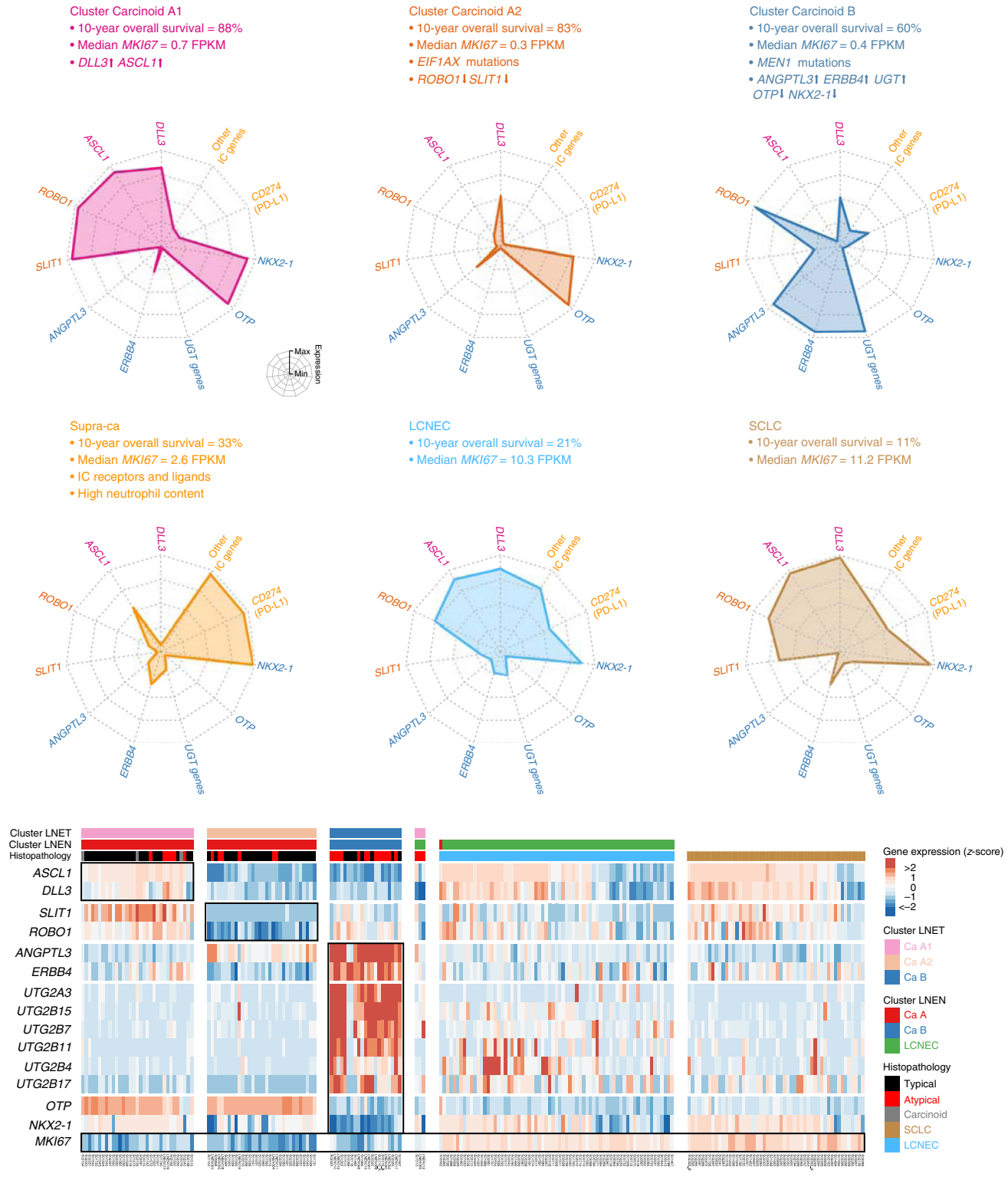


Fig. 6 Main molecular and clinical characteristics of lung neuroendocrine neoplasms. Upper panel: Radar charts of the expression level (*z*-score) of the characteristic genes [*DLL3, ASCL1, ROBO1, SLIT1, ANGPTL3, ERBB4*, UGT genes family, *OTP, NKX2-1, PD-L1 (CD274)*, and other immune checkpoint genes] of each LNET molecular cluster (Carcinoid A1, Carcinoid A2, and B clusters), supra-ca, LCNEC, and SCLC. The coloured text lists relevant characteristics—additional molecular, histopathological, and clinical data—of each group. Lower panel: heatmap of the expression level (*z*-score) of the characteristic genes of each group from the left panel, expressed in *z*-scores. Data necessary to reproduce the figure are provided in the European Genome-phenome Archive

of human cancers[37]. Similarly, overexpression of the epidermal growth factor receptor *ERBB4*, which induces a variety of cellular responses, including mitogenesis and differentiation, has also been associated with several cancer types[38,39].

For many years, it has been widely accepted that the lung well-differentiated NETs (typical and atypical carcinoids) have unique clinico-histopathological traits with no apparent causative relationship or common genetic, epidemiologic, or clinical traits with

the lung poorly differentiated SCLC and LCNEC[3]. While molecular studies have sustained this belief for pulmonary carcinoids vs. SCLC[11,13,14], the identification of a carcinoid-like group of LCNECs[10,12], the recent observation of LCNEC arising within a background of pre-existing atypical carcinoid[40], and a recent proof-of-concept study supporting the progression from pulmonary carcinoids to LCNEC and SCLC[9], suggest that the separation between pulmonary carcinoids and LCNEC might be more subtle than initially thought, at least for a subset of patients. Our study supports the suggested molecular link between pulmonary carcinoids and LCNEC, as we have identified a subgroup of atypical carcinoids, named supra-carcinoids, with a clear carcinoid morphological pattern but with molecular characteristics similar to LCNEC. In our series, the proportion of supra-carcinoids was in the order of 5.5% (six out of 110 pulmonary carcinoids with available expression/methylation data); however, considering the intermediate phenotypes observed in the MOFA LNEN, the exact proportion would need to be confirmed in larger series. We found high estimated levels of neutrophil infiltration in the supra-carcinoids. For both supra-carcinoids and LCNEC (but not SCLC), the pathways related to neutrophil chemotaxis and degranulation, were also altered. Neutrophil infiltration may act as immunosuppressive cells, for example through PD-L1 expression[41]. Indeed, the supra-carcinoids also presented levels of immune checkpoint receptors and ligands (including *PDL1* and *CTLA4*) similar—or higher—than those of LCNEC and SCLC, as well as upregulation of other immunosuppressive genes such as HLA-G, and interferon gamma that is speculated to promote cancer immune-evasion in immunosuppressive environments[42,43]. If confirmed, this would point to a therapeutic opportunity for these tumours since strategies aiming at decreasing migration of neutrophils to tumoral areas, or decreasing the amount of neutrophils have shown efficacy in preclinical models[44]. Similarly, immune checkpoint inhibitors, currently being tested in clinical trials, might also be a therapeutic option for these patients.

Overall, although preliminary, our data suggest that supra-carcinoids could be diagnosed based on a combination of morphological features (carcinoid-like morphology, useful for the differential diagnosis with LCNEC/SCLC) and the high expression of a panel of immune checkpoint (IC) genes (LCNEC/SCLC-like molecular features, useful for the differential diagnosis with other carcinoids); the levels of IC genes, such as *PD-L1*, *VISTA*, and *LAG3*, could also be used to drive the therapeutic decision for patients harbouring a tumour belonging to this subset of very aggressive carcinoids. Nevertheless, due to the very low number of supra-carcinoids identified so far ($n = 6$), follow-up studies are warranted to comprehensively characterise these tumours from pathological and molecular standpoints, to evaluate the immune cell distribution, and to establish if the diagnosis of these supra-carcinoids can be undertaken in small biopsies. Finally, the current classification only recognises the existence of grade-1 (typical) and grade-2 (atypical) well-differentiated lung NETs, while the grade-3 would only be associated with the poorly differentiated SCLC and LCNEC; however, in the pancreas, stomach and colon, the group of well-differentiated grade-3 NETs are well known and broadly recognised[45]. Whether these supra-carcinoids constitute a separate entity that may be the equivalent in the lung of the gastroenteropancreatic, well-differentiated, grade-3 NETs will require further research.

In summary, this study provides comprehensive insights into the molecular characteristics of pulmonary carcinoids, especially of the understudied atypical carcinoids. We have identified three well-characterised molecular groups of pulmonary carcinoids with different prognoses and clinical implications. Finally, the identification of supra-carcinoids further supports the already suggested molecular link between pulmonary carcinoids and LCNEC that warrants further investigation.

Methods

Sample collection. All new specimens were collected from surgically resected tumours, applying local regulations and rules at the collecting site, and including patient consent for molecular analyses as well as collection of de-identified data, with approval of the IARC Ethics Committee. These samples underwent an independent pathological review. For the typical carcinoids and LCNEC, on which methylation analyses were performed, the DNA came from the samples included in already published studies[4,11–14,35], for which the pathological review had already been done.

Clinical data. Collected clinical data included age (in years), sex (male or female), smoking status (never smoker, former smoker, passive smoker, and current smoker), Union for International Cancer Control/American Joint Committee on Cancer stage, professional exposure, and survival (calculated in months from surgery to last day of follow-up or death). These data were merged with that from Fernandez-Cuesta et al.[11], George et al.[12], and George et al.[14]. In order to improve the power of the statistical analyses, we regrouped some levels of variables that had few samples. Age was discretized into three categories ((15, 40], (40, 60], and (60, 90] years), Union for International Cancer Control stages were regrouped into four categories (I, II, III, IV), and smoking status was regrouped into two categories (non-smoker, that includes never smokers and passive smokers, and smoker, that includes current and former smokers). In addition, one patient (S02236) that was originally classified as male was switched to female based on its concordant whole-exome, transcriptome, and methylome data; and one patient (LNEN028) for whom no sex information was available was classified as male based on its methylation data (Supplementary Fig. 28; see details of the methods used in the DNA sequencing, expression, and methylation sections of the methods), because we had no other data type for this sample. Note that two SCLC samples from George et al.[14] displayed Y chromosome expression patterns discordant with their clinical data (S02249 and S02293; Supplementary Fig. 28B), but because we did not perform any analysis of SCLC samples that used sex information, this did not have any impact on our analyses. See Supplementary Data 1 for the clinical data associated with the samples.

We assessed the associations between clinical variables—a batch variable (sample provider), the main variable of interest (histopathological type), and important biological covariables (sex, age, smoking status, and tumour stage)—using Fisher's exact test, adjusting the *p*-values for multiple testing. Using samples from all histopathological types (typical and atypical carcinoids, LCNEC, and SCLC), we found that the sample provider was significantly associated with the histopathological type (Supplementary Fig. 29A). Indeed, the 20 carcinoids from one of the providers (provider 1) are all atypical carcinoids. Nevertheless, because there are also seven atypical carcinoids from a second provider and five from a third one, variables provider and histopathological type are not completely confounded and we could check for batch effects in the following molecular analysis by making sure that the molecular profiles of atypical carcinoids from provider 1 overlap with that from the two other providers. The histopathological type was significantly associated with all other variables (Supplementary Fig. 29A, B, and C).

Pathological review. Some of the samples included in this manuscript had already undergone a Central Pathological Review in the context of other published studies, so we used the classifications from the supplementary tables of the corresponding manuscripts[4,11,12,14,35]. For the new ones, an H&E (hematoxylin and eosin) stain from a representative FFPE block was collected for all tumours for pathological review. All tumours were classified according to the 2015 WHO classification by three independent pathologists (E.B., B.A.A., and S.L.). An H&E stain was also performed in order to assess the quality of the frozen material used for molecular analyses and to confirm that all frozen samples contained at least 70% of tumour cells.

Immunohistochemistry. FFPE tissue sections (3 µm thick) from 20 atypical and typical carcinoids were deparaffinized and stained with the Ventana DLL3 (SP347) assay, UltraView Universal DAB Detection Kit (Ventana Medical Systems and Amplification Kit (Ventana Medical Systems—Roche) on Ventana ULTRA autostainer (Ventana, Roche, Meylan, France), and with the CD1 rabbit monoclonal antibody (cl EP3622) (Ventana). The positivity of DLL3 was defined by the percentage of tumour cells exhibiting a cytoplasmic staining, whatever the intensity. The positivity of CD1A was defined by the percentage of the total surface of the tumour exhibiting a membrane staining with 1 corresponding to less than 1%, 2 to a percentage between 1 and 5%, and 3 to greater than 5%. Results are presented in Supplementary Data 9 and representative slides are shown in Fig. 4c.

Statistical analyses. All tests involving multiple comparisons were adjusted using the Benjamini–Hochberg procedure controlling the false discovery rate[46] using the p.adjust R function (stats package version 3.4.4). All tests were two-sided. Also, a

summary of the statistics associated with survival analyses is provided in Supplementary Data 14.

Survival analysis. We performed survival analysis using Cox's proportional hazard model; we assessed the significance of the hazard ratio between the reference and the other levels using Wald tests, and assessed the global significance of the model using the logrank test statistic (R package survival v. 2.41-3). Kaplan–Meier and forest plots were drawn using R package survminer (v. 0.4.2). Note that three LCNEC samples from George et al.[14] had missing survival censor information and were thus excluded from the analysis (samples S01580, S01581, and S01586).

DNA extraction. Samples included were extracted using the Gentra Puregene tissue kit 4g (Qiagen, Hilden, Germany), following the manufacturer's instructions. All DNA samples were quantified by the fluorometric method (Quant-iT PicoGreen dsDNA Assay, Life Technologies, CA, USA), and assessed for purity by NanoDrop (Thermo Scientific, MA, USA) 260/280 and 260/230 ratio measurements. DNA integrity of Fresh Frozen samples was checked by electrophoresis in a 1.3% agarose gel.

RNA extraction. Samples included were extracted using the Allprep DNA/RNA extraction kit (Qiagen, Hilden, Germany), following manufacturer's instructions. All RNA samples were treated with DNAse I for 15 min at 30 °C. RNA integrity of frozen samples was checked with Agilent 2100 Electrophoresis Bioanalyser system (Agilent Biotechnologies, Santa Clara, CA95051, United States) using RNA 6000 Nano Kit (Agilent Biotechnologies).

Whole-genome sequencing (WGS). Whole-genome sequencing was performed on three fresh frozen pulmonary carcinoids and matched-blood samples by the Centre National de Recherche en Génomique Humaine (CNRGH, Institut de Biologie François Jacob, CEA, Evry, France). After a complete quality control, genomic DNA (1 µg) has been used to prepare a library for whole-genome sequencing, using the Illumina TruSeq DNA PCR-Free Library Preparation Kit (Illumina Inc., CA, USA), according to the manufacturer's instructions. After normalisation and quality control, qualified libraries have been sequenced on a HiSeqX5 platform from Illumina (Illumina Inc., CA, USA), as paired-end 150 bp reads. One lane of HiSeqX5 flow cell has been produced for each sample, in order to reach an average sequencing depth of 30x for each sample. Sequence quality parameters have been assessed throughout the sequencing run and standard bioinformatics analysis of sequencing data was based on the Illumina pipeline to generate fatsq files for each sample.

Whole-exome sequencing (WES). Whole-exome sequencing was performed on 16 fresh frozen atypical carcinoids in the Cologne Centre for Genomics. Exomes were prepared by fragmenting 1 µg of DNA using sonication technology (Bioruptor, Diagenode, Liège, Belgium) followed by end repair and adapter ligation including incorporation of Illumina TruSeq index barcodes on a Biomek FX laboratory automation workstation from Beckman Coulter (Beckman Coulter, Brea, CA, USA). After size selection and quantification, pools of five libraries each were subjected to enrichment using the SeqCap EZ v2 Library kit from NimbleGen (44Mb). After validation (2200 TapeStation; Agilent Technologies, CA, USA), the pools were quantified using the KAPA Library Quantification kit (Peqlab, Erlangen, Germany) and the 7900HT Sequence Detection System (Applied Biosystems, Waltham, MA, USA), and subsequently sequenced on an Illumina HiSeq 2000 sequencing instrument using a paired-end 2 × 100 bp protocol and an allocation of one pool with 5 exomes/lane. The expected average coverage was approximately 120x after removal of duplicates (11 GB).

Targeted sequencing. Targeted sequencing was performed on the same 16 fresh frozen atypical carcinoids and 13 matched-normal tissue for the samples with enough DNA. Three sets of primers covering 1331 amplicons of 150–200 bp were designed with the QIAGEN GeneRead DNAseq custom V2 Builder tool on GRCh37 (gencode version 19). Target enrichment was performed using the GeneRead DNAseq Panel PCR Kit V2 (QIAGEN) following a validated in-house protocol (IARC). The multiplex PCR was performed with six separated primers pools [(1) 1 pool covering 786 amplicons, (2) 4 pools covering 498 amplicons, and (3) 1 pool covering 47 amplicons]. Per pool, 20 ng (1) or 10 ng (2 and 3) of DNA were dispensed and air-dried (only 2 and 3). Subsequently 11 µL (1) or 5 µL (2 and 3) of the PCR mix were added [containing 5.5 µL (1) or 2.5 µL (2 and 3) Primer mix pool (2x), 2.2 µL (1) or 1 µL (2 and 3) PCR Buffer (5x), 0.73 µL (1) or 0.34 µL (2 and 3) HotStar Taq DNA Polymerase (6 U/µL) and 0.57 µL (1) or 1.16 µL (2 and 3) H2O] and the DNA were amplified in a 96-well-plate as following: 15 min at 95 °C; 25 (1), 21 (2), or 23 (3) cycles of 15 s at 95 °C and 4 min at 60 °C; and 10 min at 72 °C. For each sample, amplified PCR products were pooled together, purified using 1.8x volume of SeraPure magnetic beads (prepared in-house following protocol developed by Faircloth & Glenn, Ecol. And Evol. Biology, Univ. of California, Los Angeles) (1) or NucleoMag® NGS Clean-up from Macherey-Nagel (2 and 3) and quantified by Qubit DNA high-sensitivity assay kit (Invitrogen Corporation). One-hundred nanograms of purified PCR product (6 µL) were used for the library preparation with the NEBNext Fast DNA Library Prep Set (New England BioLabs) following an in-house validated protocol (IARC). End repair was performed [1.5 µL of NEBNext End Repair Reaction Buffer, 0.75 µL of NEBNext End Repair Enzyme Mix, and 6.75 µL of H2O] followed by ligation to specific adapters and in-house prepared individual barcodes (Eurofins MWG Operon, Germany) [4.35 µL of H2O, 2.5 µL of T4 DNA Ligase Buffer for Ion Torrent, 0.7 µL of Ion P1 adaptor (double-stranded), 0.25 µL of Bst 2.0 WarmStart DNA Polymerase, 1.5 µL of T4 DNA ligase, and 0.7 µL of in-house barcodes]. Bead purification of 1.8x was applied to clean libraries and 100 ng of adaptator ligated DNA were amplified with 15 µL of Master Mix Amplification [containing 1 µL of Primers, 12.5 µL of NEBNext High-Fidelity 2x PCR Master Mix, and 1.5 µL of H2O]. Pooling of libraries was performed equimolarly and loaded on a 2% agarose gel for electrophoresis (220 V, 40 min). Using the GeneClean™ Turbo kit (MP Biomedicals, USA) pooled DNA libraries were recovered from selected fragments of 200–300 bp in length. Libraries quality and quantity were assessed using Agilent High Sensitivity DNA kit on the Agilent 2100 Bioanalyzer on-chip electrophoreses (Agilent Technologies). Sequencing of the libraries was performed on the Ion Torrent™ Proton Sequencer (Life Technologies Corp) aiming for deep coverage (> 250x), using the Ion PI™ Hi-QT™ OT2 200 Kit and the Ion PI™ Hi-Q™ Sequencing 200 Kit with the Ion PI™ Chip Kit v3 following the manufacturer's protocols.

DNA data processing. WGS and WES reads mapping on reference genome GRCh37 (gencode version 19) were performed using our in-house workflow (https://github.com/IARCbioinfo/alignment-nf, revision number 9092214665). This workflow is based on the nextflow domain-specific language[47] and consists of three steps: reads mapping (software bwa version 0.7.12-r1044)[48], duplicate marking (software samblaster, version 0.1.22)[49], and reads sorting (software sambamba, version 0.5.9)[50]. Reads mapping for the targeted sequencing data was performed using the Torrent Suite software version 4.4.2 on reference genome hg19. Local realignment around indels was then performed for both using software ABRA (version 0.97bLE)[51] on the regions from the bed files provided by Agilent (SeqCap_EZ_Exome_v2_probe-covered.bed) and QIAGEN, respectively, for the WES and targeted sequencing data. Consistency between sex reported in the clinical data and WES data was assessed by computing the total coverage on X and Y chromosomes (Supplementary Fig. 28A).

Variant calling and filtering on DNA. *WES data*: We re-performed variant calling for all typical and atypical carcinoid WES, including already published data, in order to remove the possible cofounding effect of variant calling in the subsequent molecular characterisation of carcinoids. Software Needlestack v1.1 (https://github.com/IARCbioinfo/needlestack)[52] was used to call variants. Needlestack is an ultra-sensitive multi-sample variant caller that uses the joint information from multiple samples to disentangle true variants from sequencing errors. We performed two separate multi-sample variant callings to avoid technical batch effects: (1) The 16 WES atypical carcinoids newly sequenced in this study were analysed together with 64 additional WES samples sequenced using the same protocol from another study in order to increase the accuracy of Needlestack to estimate the sequencing error rate; (2) The 15 WES LNET (ten typical and five atypical carcinoids) previously analysed (Fernandez-Cuesta et al.)[11] were reanalysed with their matched-normal. For both variant callings, we used default software parameters except for the minimum median coverage to consider a site for calling, the minimum mapping quality, and the SNV and INDEL strand bias[13] threshold (they were set to 20, 13, 4, and 10, respectively). Annotation of resulting variant calling format (VCF) files was then performed with ANNOVAR (2018Aprl16)[53] using the PopFreqAll (maximum frequency over all populations in ESP6500, 1000G, and ExAC germline databases), COSMIC v84, MCAP, REVEL, SIFT, and Polyphen (dbnsfp30a) databases.

We performed the same variant filtering after each of the two variant callings, based on several stringent criteria. First, we only retained variants that have never been observed in germline databases or present at low frequency (≤ 0.001) but already reported as somatic in the COSMIC database. Second, we only retained variants that were in coding regions and that had an impact on expressed proteins: we filtered out silent, non-damaging single nucleotide variants (based on MCAP, REVEL, SIFT, or Polyphen2 databases) and variants present in non-expressed genes (mean and median FPKM < 0.1 over all carcinoid tumours). Additionally, for calling (2), we re-assessed the somatic status of variants reported by Needlestack in light of possible contamination errors. Indeed, Needlestack is a very sensitive caller and will sometimes detect low allelic fraction variants in normal tissue that actually come from contamination by tumour cells. In such cases the variant is found in both matched samples and is reported as germline, but we still considered a variant as somatic if its allelic fraction in the normal tissue was at least five times lower than the allelic fraction observed in the tumour.

Targeted sequencing data: Software Needlestack was also used to call variants on targeted sequencing data from 16 atypical carcinoids and their matched-normal tissue. We performed the calling with default parameters except for the phred-scaled q-value and minimum median coverage to consider a site (20 and 10, respectively). These parameters were decreased compared to WES variants calling because we wanted a larger sensitivity in the validation set than in the discovery set. The annotation procedure was the same as for WES data. No other filters were used.

Validation: For both previously published data and data generated in this study, we only report somatic mutations that were validated using a different technique: targeted sequencing, RNA sequencing (see below for variant calling in RNA-seq data), or Sanger sequencing. Results are presented Supplementary Data 4.

Structural variant calling. Somatic copy number variations (CNVs) were called from WGS data using an in-house pipeline (software WGinR, available at https://github.com/aviari/wginr) that consists of three main steps. First, the dependency between GC content and raw read count is modelled using a generalised additive smoothing model with two nested windows in order to catch short and long distance dependencies. The model is computed on a subset of human genome mappable regions defined by a narrow band around the mode of binned raw counts distribution. This limits the incorporation of true biological signal (losses and gains) by selecting only regions with (supposedly) the same ploidy. In a second step, we collect heterozygous positions in the matched-normal sample and GC-corrected read counts (RC) and alleles frequencies (AF) at these positions are used to estimate the mean tumour ploidy and its contamination by normal tissue. This ploidy model is then used to infer the theoretical absolute copy number levels in the tumour sample. In the third step, a simultaneous segmentation of RC and AF signals (computed on all mappable regions) is performed using a bivariate Hidden Markov Model to generate an absolute copy number and a genotype estimate for each segment.

Somatic structural variants (SV) were identified using an in-house tool (crisscross, available at https://github.com/anso-sertier/crisscross) that uses WGS data and two complementary signals from the read alignments: (a) discordant pair mapping (wrong read orientation or incorrect insert-size) and (b) soft-clipping (unmapped first or last bases of reads) that allows resolving SV breakpoints at the base pair resolution. A cluster of discordant pairs and one or two clusters of soft-clipped reads defined an SV candidate: the discordant pairs cluster defined two associated regions, possibly on different chromosomes and the soft-clipped reads cluster(s), located in these regions, pinpointed the potential SV breakpoint positions. We further checked that the soft-clipped bases at each SV breakpoint were correctly aligned in the neighbourhood of the associated region. SV events were then classified as germline or somatic depending on their presence in the matched-normal sample. Results are presented as Supplementary Data 8 and one sample is highlighted in Fig. 3c.

Gene-set enrichment analysis of somatic mutations. Gene-set enrichment for somatic mutations was assessed independently for each set of Hallmark of cancer genes[18] using Fisher's exact test. We built the contingency tables used as input of the test taking into account genes with multiple mutations and used the fisher.test R function (stats package version 3.4.4). We also included validated mutations (we removed silent and intron/exon mutations) reported in SCLC[13]. In each group the *p*-values given by Fisher's exact test performed for all Hallmarks were adjusted for multiple testing. Supplementary Data 5 lists the altered hallmarks, including the mutated genes and the associated *q*-value for each group, as well as the mutated genes for each hallmarks present in each supra-carcinoid, cluster LNET, LCNEC, and SCLC samples.

We performed several robustness analyses to assess the validity of our results, in particular with regards to outlier samples/genes that would have a high leverage on the statistical results, i.e., that would alone drive the significance of a particular hallmark. First, we assessed the leverage of each individual sample using a jackknife procedure (i.e., for each sample, we performed the GSE test after removing this sample). Second, we assessed the leverage of each gene using a jackknife procedure (i.e., for each gene, we performed the GSE test without this gene). We observed that when we removed sample LNEN010 from the cluster LNET B, the sustaining proliferative signalling hallmark enrichment became non-significant at the 0.05 false discovery rate threshold, but was still significant at the 10% threshold (q-value = 0.075; Supplementary Data 3). Similarly, we observed that for several SCLC samples, once the sample was removed, the deregulating cellular energetics and inducing angiogenesis hallmarks became significant at the 0.05 false discovery rate threshold (Supplementary Data 5). For supra-carcinoids samples, we performed GSE for each sample individually. The code used for the gene set enrichment analyses on somatic mutations (Hallmarks_of_cancer_GSEA.R) is available in the Supplementary Software file 1 and the associated results are reported in Supplementary Data 5.

RNA sequencing. RNA sequencing was performed on 20 fresh frozen atypical carcinoids in the Cologne Centre for Genomics. Libraries were prepared using the Illumina® TruSeq® RNA sample preparation Kit. Library preparation started with 1 µg total RNA. After poly-A selection (using poly-T oligo-attached magnetic beads), mRNA was purified and fragmented using divalent cations under elevated temperature. The RNA fragments underwent reverse transcription using random primers. This is followed by second strand complementary DNA (cDNA) synthesis with DNA Polymerase I and RNase H. After end repair and A-tailing, indexing adapters were ligated. The products were then purified and amplified (14 PCR cycles) to create the final cDNA libraries. After library validation and quantification (Agilent 2100 Bioanalyzer), equimolar amounts of library were pooled. The pool was quantified by using the Peqlab KAPA Library Quantification Kit and the Applied Biosystems 7900HT Sequence Detection System. The pool was sequenced by using an Illumina TruSeq PE Cluster Kit v3 and an Illumina TruSeq SBS Kit v3-HS on an Illumina HiSeq 2000 sequencer with a paired-end (101x7x101 cycles) protocol.

RNA data processing. The 210 raw reads files (89 carcinoids, 69 LCNEC, 52 SCLC) were processed in three steps using the RNA-seq processing workflow based on the nextflow language[47] and accessible at https://github.com/IARCbioinfo/RNAseq-nf (revision da7240d). (i) Reads were scanned for a part of Illumina's 13 bp adapter sequence 'AGATCGGAAGAGC' at the 3' end using Trim Galore v0.4.2 with default parameters. (ii) Reads were mapped to reference genome GRCh37 (gencode version 19) using software STAR (v2.5.2b)[54] with recommended parameters[55]. (iii) For each sample, a raw read count table with gene-level quantification for each gene of the comprehensive gencode gene annotation file (release 19, containing 57,822 genes) was generated using script htseq-count from software htseq (v0.8.0)[56]. Gene fragments per kilobase million (FPKM) of all genes from the gencode gene annotation file were computed using software StringTie (v1.3.3b)[57] in single pass mode (no new transcript discovery), using the protocols from Pertea et al.[57] (nextflow pipeline accessible at https://github.com/IARCbioinfo/RNAseq-transcript-nf; revision c5d114e42d).

Quality control of the samples was performed at each step. Software FastQC (v. 0.11.5; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check raw reads quality, software RSeQC (v. 2.6.4) was used to check alignment quality (number of mapped reads, proportion of uniquely mapped reads). Software MultiQC (v. 0.9)[58] was used to aggregate the QC results across samples. Concordance between sex reported in the clinical data and sex chromosome gene expression patterns was performed by comparing the sum of variance-stabilised read counts (vst function from R package DESeq2) of each sample on the X and Y chromosomes (Supplementary Fig. 28B).

Variant calling on RNA. Software Needlestack was also used to call variants on the 20 RNA sequencing data for WES variant validation. Default parameters were used, except for the phred-scaled *q*-value, minimum median coverage to consider a site, and minimum mapping quality (20, 10, and 13, respectively). The annotation procedure was the same as for WES data.

Fusion transcript detection. RNA-seq data was processed as previously described[11,13] to detect chimeric transcripts. In brief, paired-end RNA-seq reads were mapped to the human reference genome (NCBI37/hg19) using GSNAP. Potential chimeric fusion transcripts were identified using software TRUP[59] by discordant read pairs and by individual reads mapping to distinct chromosomal locations. The sequence context of rearranged transcripts was reconstructed around the identified breakpoint and the assembled fusion transcript was then aligned to the human reference genome to determine the genes involved in the fusion. All interesting fusion-transcript were validated by Sanger sequencing. The code used for the fusion transcript detection is available on https://github.com/ruping/TRUP. All the associated results are presented Supplementary Data 7, and selected genes are highlighted in Fig. 3b.

Unsupervised analyses of expression data. The raw read counts of 57,822 genes from the 210 samples were normalised using the variance stabilisation transform (vst function from R package DESeq2 v1.14.1)[60]; this transformation enables comparisons between samples with different library sizes and different variances in expression across genes. We removed genes from the sex-chromosomes in order to reduce the influence of sex on the expression profiles, resulting in a matrix of gene expression with 54,851 genes and 210 samples. We performed four analyses, with different subsets of samples. (i) An analysis with all 210 samples (LNEN and SCLC), (ii) an analysis with LNEN samples only (158 samples), (iii) an analysis with LNET and SCLC samples only (139 samples), and (iv) an analysis with LNET samples only (89 samples). For each analysis, the most variable genes (explaining 50% of the total variance in variance-stabilised read counts) were selected (6398, 6009, 6234, and 5490 genes, respectively, for i, ii, iii, and iv). Principal component analysis (PCA) was then performed independently for each analysis (function dudi.pca from R package ade4 v1.7-8)[61]. Results are presented in Supplementary Fig. 6; see the Multi-omic integration section of the methods for a comparison of the results of the unsupervised analysis of expression data with that of the other 'omics.

We used the results from the PCA to detect outliers and batch effects in the expression data set. We did not detect any outliers in any of the analyses from Supplementary Fig. 6. We further studied the association between expression data, batch (sample provider), and five clinical variables of interest (histopathological type, age, sex, smoking status, and stage) using a PCA regression analysis. For each principal component, we fitted separate linear models with each of the six covariables of interest (provider plus the five clinical variables) and adjusted the resulting *p*-values for multiple testing. Results highlighted an association between principal component 2 and provider, histopathological type, and sex, and an association between principal components 4 and 5 and stage (Supplementary Fig. 30A). The fact that both histopathology and sample provider are jointly significantly associated with PC2 is expected given their non-independence (Supplementary Fig. 29A, B). In order to assess whether there was a batch effect

explaining the variation on PC2, we investigated the range of samples from each provider on PC2 (Supplementary Fig. 30B). We can see that samples from Provider 1 and provider 2 span a similar range on PC2 (from values less than −20 to values greater than 40). Restricting the analysis to atypical carcinoids, we can further see that AC samples from provider 2 have a range included in that of provider 1, which is expected given their differing sample sizes (five from provider 2 compared to 20 from provider 1). Overall, this shows that samples from the two providers have similar profiles and can be combined. In addition, we found that the samples that were independently sequenced in a previous study[11] and in this study (samples S00716_A and S00716_B, respectively) were spatially close in the PCA (technical replicates highlighted in Supplementary Fig. 30B).

Supervised analysis of expression data. We performed three distinct differential expression (DE) analyses. (i) A comparison between histopathological types; (ii) A comparison between pulmonary carcinoid (LNET) clusters A1, A2, and B (see Fig. 5a and the Multi-omic integration method section); (iii) a comparison between lung neuroendocrine neoplasm (LNEN) clusters Carcinoid A, Carcinoid B, and LCNEC (see the Multi-omic integration method section).

For each differential expression (DE) analysis, among the 57,822 genes from the raw read count tables, genes that were expressed in less than 2 samples were removed from the analysis, using a threshold of 1 fragment per million reads aligned. We also removed samples with missing data in the variables of interest (either histopathological types, LNET clusters, or LNEN clusters) or in any of the clinical covariables included in the statistical model (sex and age). This resulted in excluding two samples with missing age data from the three analyses (samples S01093, S02236), and further excluding three samples with no clear histopathological type (classified as carcinoids in Supplementary Data 1) from analysis (i) (samples S00076, S02126, S02154). For each analysis, we then identified DE genes from the raw read counts using R package DESeq2 (v. 1.21.5)[60]. For each analysis, we fitted a model with the variable of interest (type, LNET cluster, or LNEN cluster) and using sex (two levels: male and female), and age (three levels: (16, 40], (40, 60], (60, 90]) as covariables. We then extracted DE genes between each pair of groups, and adjusted the p-values for multiple testing. In order to select the genes that have the largest biological effect, we tested the null hypothesis that the two focal groups had less than 2 absolute $\log_2$-fold changes differences. For each analysis, we define the core genes of a focal group as the set of genes that are DE in all pairwise comparisons between the focal group and other groups; they correspond to genes, which expression level is specific to the focal group. For example, given three groups—A, B, and C—to find core genes, which expression levels uniquely define A compared to both B and C, we select DE genes that differentiate A from B (A vs. B), DE genes that differentiate A from C (A vs. C) and take the intersection of these gene sets [(A vs. B)∩(A vs. C)]. The code used for the DE analyses (RNAseq_supervised.R) is available at https://github.com/IARCbioinfo/RNAseq_analysis_scripts. Results of analysis (i) are reported in Supplementary Data 15 and Supplementary Fig. 31; results of analysis (ii) are reported in Supplementary Data 10 and Fig. 5a; results of analysis (iii) are reported in Supplementary Data 12. See section Multi-omics integration for comparisons between the analyses based on histopathological types [analysis (i)] from all 'omics perspectives.

Note that an alternative method for finding DE genes would be to compare a focal group to all the other samples together. For example, comparing group A to both groups B and C simultaneously [denoted A vs. (B and C) or A vs. the rest]. Note that this would find genes that are DE between A and the average level of expression of B and C, and thus this alternative method would have the unwanted behaviour of including the genes that are strongly DE in the comparison of A vs. B, but with similar expression levels in A and C. In order to compare the methods we used to detect core genes with this alternative method, we performed an analysis similar to analysis (ii) but comparing a focal group to all the other samples simultaneously (A vs. the rest). The comparison between our method and the alternative one is presented in Supplementary Fig. 21 and shows that our analysis provides conservative results compared to testing the focal group vs. the rest. Indeed, core DE genes reported are almost exclusively a subset of the genes found when comparing the focal group vs. the rest.

Immune contexture deconvolution from expression data. We quantified the proportion of cells that belong to each of ten immune cell types (B cells, macrophages M1, macrophages M2, monocytes, neutrophils, NK cells, CD4+ T cells, CD8+ T cells, CD4+ regulatory T cells, and dendritic cells) from the RNA-seq data using software quanTIseq (downloaded 23 March 2018)[62]. quanTIseq uses a rigorous RNA-seq processing pipeline to quantify the gene expression of each sample, and performs supervised expression deconvolution in a set of genes identified as informative on immune cell types, using the least squares with equality/inequality constrains (LSEI) algorithm with a reference data set containing expected expression levels for the ten immune cell types. Importantly, quanTIseq also provides estimates of the total proportion of cells in the bulk sequencing that do and do not belong to immune cells.

We tested whether immune composition differed between histopathological types, LNET clusters, LNEN clusters, and supra-carcinoids using linear permutation tests (R package lmperm, v. 2.1.0). Permutations tests are exact statistical tests that do not rely on approximations and assumptions regarding the data distribution, and are thus well-fitted to test whether a few samples come from the same distribution as a larger group of samples. As such, they were well-fitted to handle the tests involving supra-carcinoids, for which only three samples had RNA-seq data. For each of the three analyses (histopathology, LNET clusters, and LNEN clusters), and for each pair of groups, we fitted one model per immune cell type, with the proportion of this cell type in each sample as explained variable and the cluster membership as explanatory variable. We adjusted the p-values for multiple testing. The code used for these three analyses is available on https://icbi.i-med.ac.at/software/quantiseq/doc/index.html and the associated results are presented Figs. 2f, 4b, and Supplementary Figs. 15, 19, and 32.

EPIC 850k methylation array. Epigenome analysis was performed on 33 typical carcinoids, 23 atypical carcinoids, and 20 LCNEC, plus 19 technical replicates. Epigenomic studies were performed at the International Agency for Research on Cancer (IARC) with the Infinium EPIC DNA methylation beadchip platform (Illumina) used for the interrogation of over 850,000 CpG sites (dinucleotides that are the main target for methylation). Each chip encompasses eight samples, so 12 chips were needed for the 95 samples. We used stratified randomisation to mitigate the batch effects, ensuring that the three histopathological types were present on every chip, while also controlling for potential confounders (the sample provider, sex, smoking status, and age of the patient); replicates were placed on different chips.

For each sample, 600 ng of purified DNA were bisulfite converted using the EZ-96 DNA Methylation-Gold™ kit (Zymo Research Corp., CA, USA) following the manufacturer's recommendations for Infinium assays. Three replicates included half the amount (300 ng). Then, 200 ng of bisulfite-converted DNA was used for hybridisation on Infinium Methylation EPIC beadarrays, following the manufacturer's protocol (Illumina Inc.). This array shares the Infinium HD chemistry (Illumina Inc.) and a similar laboratory protocol used to interrogate the cytosine markers with HumanMethylation450 beadchip. Chips were scanned using Illumina iScan to produce two-colour raw data files (IDAT format).

Methylation data processing. The resulting IDAT raw data files were pre-processed using R packages minfi (v. 1.24.0)[63] and ENmix (v. 1.14.0)[64]. We first removed unwanted technical variation in-between arrays using functional normalisation of the raw two-colour intensities, and computed the β-values for the 866,238 probes and 96 samples. Then, we filtered four types of probes that could confound the analyses. (i) We removed probes on the X and Y chromosomes, because we were interested in variation between tumours and treated sex as a confounder. (ii) We removed known cross-reactive probes—i.e., probes that co-hybridise to other chromosomes and thus cannot be reliably investigated. (iii) We removed probes that had failed in at least one sample, using a detection p-value threshold of 0.01, where p-values were computed with the detection P function from R package minfi, that compares the total signal (methylated + unmethylated) at each probe with the background signal level from non-negative control probes. (iv) We removed probes associated with common SNPs—that reflect underlying polymorphisms rather than methylation profiles—using a threshold minor allele frequency of 5% in database dbSNP build 137 (function dropLociWithSnps from minfi). (v) We removed probes putatively associated with rare SNPs by detecting and removing probes with multimodal β-value distributions (function nmode.mc from R package ENmix). Next, we removed duplicated samples, randomly choosing one sample per pair so as to minimise potential discrepancies, and we removed one sample that came from a metastatic tumour rather than a primary tumour. The final data set contained the β-values of 767,781 CpGs for 76 samples.

We performed quality controls of the raw data. Two-colour intensity data of internal control probes were inspected to check the quality of successive sample preparation steps (bisulfite conversion, hybridisation). We did not find outliers when comparing the methylated/unmethylated channel intensities of all samples, nor did we find samples with overall low detection p-values (the sample with the lowest mean p-value had a value of 0.001). Concordance between the sex reported in the clinical data and the methylation data was assessed using a predictor based on the median total intensity on sex-chromosomes, with a cutoff of -2 $\log_2$ estimated copy number (function getSex from minfi). Consistently with the WES and RNA-seq data, we found one sample with a mismatch between reported and inferred sex (see results in Supplementary Fig. 28C). We investigated batch effects at the raw data level using surrogate variable analysis. We used function ctrlsva from package ENmix to compute a principal component analysis of the intensity data from non-negative control probes. We retained the first ten principal components—hereafter referred to as surrogate variables—explaining >90% of the variation in control probes intensity. The ten surrogate variables were included as covariables in later supervised analyses to mitigate the impact of batch effects on the results. We checked the association of surrogate variables with batch (chip, position on the chip, and sample provider) and clinical variables (histopathological type, age, sex, smoking status) using PCA regression analysis, fitting separate linear models to each surrogate variable with each of the seven covariables of interest and adjusted the p-values for multiple testing. We show in Supplementary Fig. 33A that surrogate variables 1, 2, 3, and 10 are significantly associated with the chip (variable Sentrix id) or position on the chip (variable Sentrix position), while surrogate variables 4, 5, and 10 are significantly associated with the sample provider. The

code used to perform all the pre-processing procedure of these data is available at https://github.com/IARCbioinfo/Methylation_analysis_scripts.

Unsupervised analysis of methylation data. The β-values of 767,781 CpGs for 76 samples were transformed into M-values to perform unsupervised analyses; indeed, contrary to β-values, M-values theoretically range from $-\infty$ to $+\infty$ and are considered normally distributed. We performed two analyses, with different subsets of samples: (i) an analysis with all carcinoid and LCNEC samples (76 samples), and (ii) an analysis with carcinoid samples only (56 samples). For each analysis, the most variable CpGs (explaining 5% of the total variance in M-values) were selected (8,483 and 7,693 CpGs, respectively, for (i) and (ii). PCA was then performed independently for each analysis (function dudi.pca from R package ade4 v1.7-8)[61]. Results are presented in Supplementary Fig. 7; see the Multi-omic integration section of the methods for a comparison of the results of the unsupervised analysis of methylation data with that of the other 'omics.

We used the results from the PCA to detect outliers and batch effects in the methylation data set. We did not detect any outliers in any of the analyses from Supplementary Fig. 7. We also performed a PCA regression analysis using the same protocol as described in the data processing section above. Results highlighted no association between any principal component and array batches (chip and position in the chip; Supplementary Fig. 33A). Principal component 2 was associated with the sample provider; further examination of the PCA (Supplementary Fig. 33B) revealed that this effect was driven by the samples from provider 1, which have the largest range of coordinates on PC2 (from < −30 to >100). Nevertheless, the fact that their coordinates on PC2 overlap with that of samples from other providers, and the fact that the vast majority of atypical carcinoid samples come from one provider, suggest that the large range of values of provider 1 samples on PC2 is driven by the biological variability of carcinoid methylation profiles. In addition, note that samples that cluster with LCNEC are not solely from provider 1. We assessed the impact of functional normalisation on batch effects by performing the same analysis on the M-values of the 5% most variable CpGs obtained without normalisation (Supplementary Fig. 33A). Compared to the PCA of the 5% most variable CpGs with normalisation (Supplementary Fig. 33A), we find that the chip position (variable Sentrix position) is significantly associated with PC10, and that PC2 is not associated with histopathology. This suggests that the functional normalisation reduced batch effects and revealed some of the biological variability in methylation data.

The PCA is also informative about associations between methylation profiles and clinical variables. We find a significant association between PC1, histopathological type, age, and smoking status, with LCNEC, smokers, and larger age classes located at higher PC1 coordinates (Supplementary Fig. 33A); these associations are expected, given that the difference between LCNEC and carcinoids is expected to be the main driver of variation in methylation, and given known the aetiology of the diseases[8]. We find an association between principal component 2, histopathology, and sex, with male and atypical carcinoids having overall larger PC2 coordinates. We find associations of larger components, in particular PC3 and age, and PC7 and 9, and sex.

Supervised analysis of methylation data. We detected differential methylation at the probe level (DMP) in three independent analyses: (i) between histopathological types (TC, AC, and LCNEC), (ii) between LNET clusters (clusters A1, A2, and B), and (iii) between LNEN clusters (clusters A, B, and LCNEC).

To detect DMPs, for each analysis, linear models were first fitted independently for each CpG to its M-values (function lmFit from R package limma version 3.34.9)[65], using the variable of interest (histopathology, LNET cluster, or LNEN cluster), in addition to the sex, age group, and the ten surrogate variables as covariables. Then, moderated t-tests were performed by empirical Bayes moderation of the standard errors (function eBayes from package limma), and p-values were computed for each CpG. Moderation enables to increase the statistical power of the test by increasing the effective degrees of freedom of the statistics, while also reducing the false-positive rate by protecting against hypervariable CpGs, and are thus favoured in array analyses. The p-values were adjusted for multiple testing, and CpGs with a q-value <0.05 were retained. The code used for the DMPs identification (DMP.R) is available in the Supplementary Software 1 and the associated results of analyses (i), (ii), and (iii) are presented Supplementary Data 16, Supplementary Data 11, and 17, respectively. See section Multi-omics integration for comparisons between the analyses based on histopathological types [analysis (i)] from all 'omics perspectives. Analysis (iii) confirmed most DMPs associated with DEGs reported in Fig. 5a for cluster B relative to LNET clusters (*TFF1*, *OTOP3*, *SLC35D3*, *APOBEC2*) were also DMPs for cluster B relative to LNEN clusters, showing that they harboured specific methylation levels that made them different from the LCNEC cluster, as well as from other carcinoid clusters.

Multi-omics integration. We performed an integrative analysis of the WES, WGS, RNA-seq, and 850 K methylation array data, using the validated somatic mutations (Supplementary Data 4), the variance-stabilised read counts, and the M-values, respectively. The full data set consisted of 243 samples, but some analyses focused on a subset of the data.

Unsupervised continuous multi-omic analyses. To perform continuous latent factors identification, we performed an integrative group factor analysis of the expression and methylation data using software MOFA (R package MOFAtools v. 0.99)[15]. MOFA identifies latent factors (LF, i.e., continuous variables) that explain most variation in the joint data sets. We did not include the somatic mutations in the model because the low level of recurrence (only four recurrently mutated genes in Supplementary Data 4) resulted in a sample by mutation matrix of much lower dimension than the other 'omics, which is known to bias the analyses[15]. Also, we did not consider expression and methylation from the sex-chromosomes, because we were interested in differences between tumours independently of the sex of the patient.

We performed four analyses, with different subsets of samples. (i) An analysis with all 235 samples for which expression or methylation data was available (LNEN and SCLC), (ii) an analysis with LNEN samples only (183 samples), (iii) an analysis with LNET and SCLC samples only (163 samples), and (iv) an analysis with LNET samples only (111 samples). For each analysis, the most variable genes for expression (explaining 50% of the total variance) were selected (6398, 6009, 6234, and 5490 genes, respectively, for i, ii, iii, and iv), and the most variable CpGs (explaining 5% of the total variance) were selected (8483, 8483, 7693, and 7693 CpGs, respectively, for i, ii, iii, and iv). Note that these lists of genes and CpGs are the same as the ones used to perform the unsupervised analyses of expression and methylation data (see above sections). Also note that we did not have EPIC 850k methylation array data for SCLC; MOFA was shown to handle missing data, including samples with entire 'omic techniques missing, by using the correlated signals from several data sets (e.g., expression and methylation) to accurately reconstruct latent factors. MOFA was performed independently for each analysis, setting the number of latent factors to 5, because subsequent latent factors explained <2% of the variance of both 'omic data sets (function runMOFA from R package MOFAtools v0.99.0). Because MOFA uses a heuristic algorithm, we assessed the robustness of the results using 20 MOFA runs. We then computed the correlations between each of the five first-latent factors across each run, resulting in a correlation matrix of 100 by 100 entries (Supplementary Figs. 2 and 17). We found that the correlations across runs were very high (> 0.95 for >80% of runs) in all analyses, suggesting that the results are robust. In addition, we found that correlations between latent factors within runs were small (typically below 0.2), which suggests that latent factors capture quasi-independent sources of variation in the data sets. For each analysis, we selected the MOFA run that resulted in the best convergence, based on the evidence lower bound statistic (ELBO). Results are presented in Figs. 1a, 4a, and Supplementary Fig. 13. Interestingly, we find that MOFA latent factors 1 to 3 for analysis (i) (LNET, LCNEC, and SCLC) correspond to MOFA LF2 to 4 for analysis (ii) (LNET and LCNEC), and to MOFA LF3 to 5 for analysis (iv) (LNET alone); this suggests that each histopathological type introduces an independent source of variation, resulting in a new LF. The code used for the unsupervised continuous molecular analyses (integration_MOFA.R) is available on https://github.com/IARCbioinfo/integration_analysis_scripts.

To perform comparisons with uni-omic unsupervised analyses, we compared the results of MOFA with that of the unsupervised analysis of expression and methylation data (Supplementary Fig. 3). To do so, we used the 51 LNEN samples for which we had both expression and methylation data, and extracted their coordinates in MOFA, expression PCA (see section unsupervised analysis of expression data), and methylation PCA (see section unsupervised analysis of methylation data). When using LNET and LCNEC samples (Supplementary Fig. 3A), we found that MOFA LF1 is strongly correlated with expression PC1 and methylation PC1 ($|r| > 0.98$; Supplementary Fig. 3D, E), and that expression PC1 and methylation PC1 are strongly correlated between them ($r = 0.97$; Supplementary Fig. 3C); LF2 was strongly correlated with expression PC3 ($r = -0.86$; Supplementary Fig. 3P), and methylation PC2 ($r = -0.98$; Supplementary Fig. 3K), suggesting that LF2 is more driven by methylation differences, but that it is nonetheless consistent with a large proportion of expression variation. On the contrary, LF3 was more strongly correlated with expression PC2 ($r = 0.87$; Supplementary Fig. 3J), suggesting that PC3 is more driven by expression differences. All these observations are consistent with the fact that the percentage of variance explained by LF2 and LF3 in terms of expression and in terms of methylation are different: LF2 explains more expression in methylation, while LF3 explains more variation in expression (Fig. 1a); it is also coherent with the fact that clusters A1 and A2 are the most separated clusters on expression PC2 (Supplementary Fig. 6B), while clusters A1 and B are the most separated on methylation PC2 (Supplementary Fig. 7A). When using LNET samples only (Supplementary Fig. 3B), we found that MOFA LF1 is strongly correlated with expression PC2 and methylation PC1 ($|r| > 0.86$; Supplementary Fig. 3M, H), and that expression PC2 and methylation PC1 are strongly correlated between them ($r = 0.72$; Supplementary Fig. 3F); LF2 was strongly correlated with expression PC1 ($r = -0.88$; Supplementary Fig. 3G), and methylation PC2 ($r = 0.90$; Supplementary Fig. 3N), suggesting that LF2 is more driven by methylation differences, but that it is nonetheless consistent with a large proportion of expression variation. Again, all these observations are consistent with the fact that the percentage of variance explained by LF1 and LF2 in terms of expression and in terms of methylation are different (Fig. 4a); it is also coherent with the fact that clusters A1 and A2 are the most separated clusters on expression PC1 (Supplementary Fig. 6D), while clusters A1 and B are the most separated on methylation PC2 (Supplementary Fig. 7B).

To perform associations of latent factors with other variables, we used the results from MOFA to detect outliers and batch effects in the data set. We did not

detect any outliers in any of the analyses from Supplementary Fig. 13. We further studied the associations between the first 5 LFs, batch (sample provider), and five clinical variables of interest (histopathological type, age, sex, smoking status, and stage) using regression analysis. For each latent factor, we fitted a linear model with the six covariables of interest (provider plus the five clinical variables). Because of the reported association between sex, age, and smoking status, we also included in the model the interaction between sex and smoking status and between age and smoking status; we adjusted the resulting p-values for multiple testing. Significant associations (q-value < 0.05) are highlighted in Figs. 1a and 4a.

We also tested the association between MOFA clusters and mutations using regression analysis. We tested genes recurrently mutated in carcinoids, using a threshold of three samples (following Argelaguet et al.)[15]; indeed, non-recurrent genes are not informative about molecular groups. Only two genes were retained: *MEN1* and *EIF1AX*. We also included recurrently mutated genes reported in LCNEC[12]. Results are highlighted in Fig. 4a. Similarly, we tested the association between pathways highlighted in Supplementary Fig. 16 (Lysine demethyltransferases, polycomb complex, SWI/SNF complex) and MOFA LF using regression analysis, but did not find any significant association at a false discovery rate threshold of 0.05.

Unsupervised discrete multi-omic analyses. We identified molecular clusters—groups of samples with similar molecular profiles—from MOFA results. Following Mo et al.[66], given a specified number of clusters K, we used the $K-1$ latent factors that explained most of the variation to perform clustering; this choice of number of latent factors in Mo et al.[66] is said to be primarily motivated by "a general principle for separating g clusters among the n datapoints, a rank-k approximation where $k \leq g-1$ is sufficient." In addition, because the MOFA latent factors explaining the most variance in gene expression and methylation are expected to capture more biological signal compared to the ones explaining the least variance—expected to represent more of the noise in the data set—we expect that using the first $K-1$ latent factors would provide more biologically meaningful clusters than using all latent factors. In addition, following the procedure from Wilkerson and Hayes[67], we performed consensus clustering to detect robust molecular clusters. This procedure involved multiple replicate clusterings (K-means algorithm; R function kmeans), each on latent factors from an independent MOFA run done on a subsample (80%) of the data. Pairwise consensus values were defined as the proportion of runs in which two samples are clustered together and used as a similarity measure, and used to perform a final hierarchical clustering (median linkage method). Consensus clustering results for K from 2 to 5, for LNET plus LCNEC samples, and LNET samples alone, are presented in Supplementary Figs. 5 and 18, respectively. In the case of LNET alone, because the optimal Dunn index, which evaluates the quality of clustering as a ratio of within-cluster to between-cluster distances, corresponded to $K=3$ clusters (Supplementary Fig. 18C), we chose the solution with three clusters. Nevertheless, note that the cluster memberships for $K=4$ and $K=5$ are almost perfectly nested into that for $K=3$ (e.g., samples from the blue cluster for $K=3$, Supplementary Fig. 18B are split between a blue and a purple cluster for $K=4$), so the solutions with three and four clusters are coherent. Cluster memberships are highlighted in Fig. 4a. Similarly, in the case of LNET plus LCNEC samples (LNEN), because the optimal Dunn index is reached when $K=3$, we chose that solution, but note that the cluster memberships for $K>3$ are also nested into that for $K=3$, so all results are coherent across values of K.

In order to test whether using additional latent factors could increase the power to detect molecular clusters, we performed a similar analysis but using all five latent factors identified by MOFA. In order to provide more importance to the factors most likely to capture the biological variation in the data, the multiple replicate clusterings were performed using a weighted k-means algorithm, where variables (here MOFA latent factors) are given weights corresponding to their proportion of variance explained. More specifically, instead of minimising the within-cluster sum of squares, the weighted within-cluster sum of squares is minimised. Results for $K=3$ clusters of LNET and LNEN samples are presented in Supplementary Fig. 8. We can see that the alternative approach (weighted K-means on five latent factors) leads to the exact same cluster membership as the original approach (K-means on $K-1$ latent factors), both for LNEN and LNET clusters. Indeed, among the latent factors, only the first 3 were associated with either the LNEN clusters (ANOVA $q = 4.09 \times 10^{-84}$, 8.63×10^{-80}, 0.66, 0.094, 0.24, respectively, for latent factors 1 through 5) or the LNET clusters (ANOVA $q = 5.06 \times 10^{-4}$, 5.99×10^{-47}, 5.12×10^{-46}, 0.15, 0.052, respectively), which indicates that the first three latent factors captured the differences between clusters. The code used for the clustering analyses (integration_unsupervised.R) is available at https://github.com/IARCbioinfo/integration_analysis_scripts.

GSEA on multi-omic latent factors. We performed gene set enrichment analysis (GSEA) on the latent factors identified by MOFA using the built-in function FeatureSetEnrichmentAnalysis[15]. This tests for each latent factor whether the distribution of the loadings of features (genes or CpGs) from a focal set are significantly different from the global distribution of loadings from features outside the set. We performed the analysis using two reference databases of gene sets: GO and KEGG. To retrieve the appropriate databases, for all genes from the multi-omics integration analysis, we downloaded GO terms using R package biomaRt[68], and we retrieved KEGG pathways using R package KEGGgraph (v. 1.38.0)[69]. Results are presented in Supplementary Data 6.

Expression and methylation correlation analysis. We performed correlation tests in two analyses: (i) between LNET clusters (clusters A1, A2, and B), and (ii) between LNEN clusters (clusters A, B, and LCNEC). We selected for each gene, the set of CpGs in the region −2000 to +2000 from the transcription start site (TSS) using function getnearestTSS from R package FDb.InfiniumMethylation.hg19 version 2.2.0 based on the IlluminaHumanMethylationEPICanno.ilm10b2.hg19 annotation (get Annotation function from R package minfi version 1.24.0)[63].

We performed correlation test analyses (function cor.test from R package stats version 3.5.1) using the core genes lists (Supplementary Data 10 and 12) to find associations between expression and methylation data for each CpG, using Pearson's correlation coefficient. The p-values were adjusted for multiple testing. In addition, we explored the correlation between expression and methylation data by fitting a linear model independently for each correlated CpG (function lm from R package stats version 3.5.1). Finally, we calculated the interquartile distance of β-values for each CpG. CpGs with a q-value < 0.05, $r^2 > 0.5$ and an interquartile distance greater than 0.25 were retained and, among these CpGs, only the one with the smallest q-value has been represented in Supplementary Fig. 22. Results of analyses (i) and (ii) are reported in Supplementary Data 10 and 12.

Survival analysis using penalised generalised linear model. We computed a generalised linear model with elastic net regularisation (R package glmnet v2.0-16)[70] to select the genes associated with the survival of LNET samples. We fixed the elastic net mixing parameter α to 0.5 and used leave-one-out cross-validation to determine the regularisation parameter λ (cv.glmnet function from glmnet package). To be more stringent, the optimal regularisation parameter chosen was the one associated with the most regularised model with cross-validation error within one standard deviation of the minimum. In order to identify the genes associated with the poor survival of the cluster Carcinoid B, we included in the model only the expression of the core genes of this cluster defined in the MOFA considering only the LNET samples (see section Multi-omics integration). We used the normalised read counts, and centred and scaled them using R package caret (v6.0-80). The genes with non-zero estimated coefficients are listed in Supplementary Data 13. For each non-coding gene, we determined the optimal cutpoint of expression (normalised read counts) that best separates the survival outcome into two groups using the surv_cutpoint function based on the maximally selected rank statistics and available in the R package survminer (v0.4.3). The minimal proportion of samples per group was set to 10%.

Supervised multi-omic analyses. We performed supervised learning in order to classify typical and atypical carcinoids, and LCNEC based on the different 'omics data available: expression and methylation data.

Classification algorithm: Each classification was performed using a random forest algorithm (R package randomForest v4.6-14). Considering the restricted number of samples, we performed a leave-one-out cross-validation. For each run, to increase the training set size, minority classes were oversampled so that all classes reach the same number of training samples. Note that for the sample with technical replication of RNA-seq data (S00716_A and S00716_B), in order to avoid model overfitting, the two replicates were never simultaneously included in the training and test sets. Also in order to avoid overfitting, we performed normalisation and independent feature filtering within each fold, so that test samples were excluded from this step. More specifically, for the expression data, the features of the training set were first normalised using the variance stabilisation transformation (vst function from R package DESeq2 v1.22.2), then mean-centred and scaled to unit variance. Then, the variance stabilising transformation learned from the training set was applied to the test set using the dispersionFunction function from the DESeq2 package, and centreing and scaling were performed using the values learned from the training set. For the methylation data, the M values were computed using the R package minfi (v1.28.3); the features of the training set were mean-centred and scaled to unit variance, then the test sample features were centred and scaled using the values learned from the training set. For each fold of the leave-one out, the training set was used for the feature selection. Based on the training set, we selected the most variable features, representing 50% and 5% of the total variation in expression and methylation data, respectively. The code used for the machine learning analyses (ML_functions.r) is available in the Supplementary Software 1 and the associated results are reported in Supplementary Data 1.

Defining an Unclassified category: The random forest algorithm provides for each predicted sample the class probabilities. We considered a sample as unclassifiable (Unclassified category) if the ratio of the two highest probabilities was below 1.5. In fact, this threshold allowed us to identify a category of samples with intermediate molecular profiles, for which the algorithm assigns similar probabilities to the two most probable classes. Because of the small sample size, this parameter was chosen a priori and not tuned in order to avoid overfitting. In Supplementary Fig. 10, we compared the classification results when considering three different thresholds: 1 (which corresponds to no ratio and results in few unclassified samples, i.e., only discordant expression and methylation-based

predictions, see Integration of expression and methylation data below), 1.5 (which corresponds to the ratio reported in the main text), and 3 (which corresponds to a very stringent ratio resulting in more unclassified samples). Except for the size of the unclassified classes that depends on the ratio used, the confusion matrices for the three ratios were qualitatively similar, with most LCNEC samples correctly classified, a majority of typical correctly classified, and almost as many atypical classified as typical and classified as atypical. In addition, the survival analyses of the three models also led to similar conclusions, with atypical carcinoids classified as atypical by the machine learning having a survival that is not statistically significantly different from that of LCNEC samples but that is lower from both that of typical carcinoids predicted as typical carcinoids, and that of atypical predicted as typical. However, in the case of the largest ratio, the small number of atypical samples predicted in those categories did not enable the identification of two groups of atypical carcinoids with significant different overall survival ($p = 0.086$).

Number of samples and features: To classify LCNEC against atypical and typical carcinoids, 157 and 76 samples were considered using the expression and methylation data, respectively. The number of features selected in each fold of the leave-one-out are of the order of 6000 and 8000 for expression and methylation features, respectively. For the analysis based on *MKI67* only (Supplementary Fig. 31C, left panel), the only feature considered was the expression of *MKI67*.

Integration of expression and methylation data: As the random forest algorithm does not handle missing data directly, and because only 51 out of 182 LNEN samples had both expression and methylation data available (Supplementary Fig. 1), we performed random forest classification on expression and methylation separately, and merged the classification results by combining the two sets of ML predictions. Thus, the samples with both expression and methylation data were associated with two predictions. When the two predictions were discordant we applied the following rules: (i) if one prediction was Unclassified (see Defining an Unclassified category above) and the other a histopathological category, we chose the histopathological category (ii) if the two predictions were different histopathological categories, we chose the Unclassified category.

Note that fitting independent random forest models on each data set separately corresponds to maximising the number of samples (n) per model at the expense of the number of features (p), because each model relies only on the number of features in a single data set. An alternative approach is to maximise the number of features (p) by combining both data sets, at the expense of the number of samples n, because of the limited number of samples with both data types available. Indeed, for fixed n increasing p requires less parameters and leads to a higher statistical power. Nevertheless, in our case, because of missing data, increasing p by using both omics layers would drastically reduce n, restricting our sample set ($n = 157$ and $n = 76$ for expression and methylation, respectively) to the set of samples with both layers ($n = 51$, including only a single supra-carcinoid). Given the existence of very rare entities such as the supra-carcinoids, accurately capturing the diversity of molecular profiles in the training set was our priority, and thus we chose to maximise n. In addition, by maximising n, we hypothetically ensured that we would also maximise the power of the subsequent analyses based on the ML results. To confirm this hypothesis, we performed the ML analyses on the restricted set of samples, including both expression and methylation data in the same model and compared the predictions of this model to the combined predictions based on expression and methylation data separately. We found that the predictions (confusion matrix in Supplementary Fig. 9) were similar, with 43/51 samples with both data types predicted similarly in the two models. In addition, our main finding —the existence of two groups of atypical samples, which tended to have a good and bad prognosis (red and pink curves Fig. 1b)—still held, but that limited number of samples impeded the statistical analyses. In fact, none of the Cox regression tests were significant even for the groups displaying the largest differences (e.g., ML-predicted LCNEC vs. ML-predicted typical samples), and even when comparing the histological types reported by the pathologists (bottom panel Supplementary Fig. 9). This supports our hypothesis that maximising p at the expense of n leads to a decrease in power in subsequent analyses due to a smaller sample size, and comforts our initial choice.

As matrix factorisation methods such as MOFA and PCA remove correlations between features by finding latent factors that summarise them, they could presumably improve the performance of ML. Nevertheless, by providing low-dimensional approximations of the data, such techniques induce a loss of information, which could reduce the performance of the ML. To assess the balance between these beneficial and detrimental effects, we also performed ML using the MOFA factors or the principal components of the PCA analysis, using factors or components that explained at least 2% of the variance (five MOFA latent factors, six expression PCs, and five methylation PCs, respectively). These analyses are presented in Supplementary Fig. 12 and led to similar classification to the results presented in the main text Fig. 1. In addition, in the case of MOFA factors, in accordance with Fig. 1, atypical carcinoids were stratified into a group with an overall survival similar to that of the LCNEC (in red) and a group with a higher overall survival (in pink), similar to that of the typical carcinoids. When using the principal components, despite a similar trend, the difference in survival between the high- and low-survival groups was not significant. These results show that dimensionality reduction does not lead to an increased classification ability, nor does it provide a better explanation of clinical behaviour. We thus chose to represent only the results of the ML analyses based on expression and methylation data in the main text and figures.

Survival analysis based on expression and methylation data. We divided the samples into different groups based on the ML predictions. We represented the Kaplan–Meier curves of the predictions groups by selecting the groups with >10 samples and gathering the unclassified samples in the same group. Using Cox's proportional hazard model and using the logrank test statistic (R package survival v2.42-3) we compared the overall survival of LCNEC, atypical and typical samples based on the histopathological classification and based on the ML predictions (Supplementary Fig. 11A). Forest plots were drawn using R package survminer (v0.4.3). The same survival analysis was performed using the ML predictions based on *MKI67* expression only (Supplementary Fig. 11C).

Comparison between the supervised analyses of typical and atypical carcinoids. We contrasted the results of the different supervised analyses between typical and atypical carcinoids based on clinical data, specific markers (Ki67), machine learning, differential expression, and differential methylation (Supplementary Fig. 31). Survival analyses showed a significant difference between histopathological types (Supplementary Fig. 31A). Nevertheless, the machine learning classifier based on the genome-wide expression or methylation data could not properly distinguish atypical and typical carcinoids (Supplementary Fig. 31B): there were 64–83% correctly classified typical carcinoids and only 30–41% correctly classified atypical carcinoids. The differential expression analysis showed that atypical carcinoids also presented very few core differentially expressed genes (Supplementary Fig. 31C, middle panel and Supplementary Data 15) and differentially methylated positions (Supplementary Fig. 31C, right panel and Supplementary Data 17). Overall, these data suggest that the histopathological classification, although clinically meaningful, does not completely match the molecular classification.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The exome sequencing data, RNA-seq data, and methylation data have been deposited in the European Genome-phenome Archive (EGA) database, which is hosted at the EBI and the CRG, under accession number EGAS00001003699. Other data sets referenced during the study are available from the EGA website under accession numbers EGAS00001000650 (pulmonary carcinoids)[11], EGAS00001000708 (LCNEC)[12], and EGAS00001000925 (SCLC)[13,14]. All the other data supporting the findings of this study are available within the article and its supplementary information files and from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

Code availability

The code and software sources from previously published algorithms used to perform the analyses are detailed in the supplementary tables and online methods. Custom scripts are provided in the Supplementary Software 1. All sources for the software used in the manuscript are summarised in Supplementary Data 18.

Received: 7 November 2018 Accepted: 2 July 2019
Published online: 20 August 2019

References

1. Travis, W. D. et al. The 2015 World Health Organization Classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* **10**, 1243–1260 (2015).
2. Rindi, G. et al. A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. *Modern Pathol.* **31**:1770-1786 (2018).
3. Caplin, M. E. et al. Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann. Oncol.* **26**, 1604–1620 (2015).
4. Swarts, D. R. et al. Interobserver variability for the WHO classification of pulmonary carcinoids. *Am. J. Surg. Pathol.* **38**, 1429–1436 (2014).
5. Thunnissen, E. et al. The Use of immunohistochemistry improves the diagnosis of small cell lung cancer and its differential diagnosis. An international reproducibility study in a demanding set of cases. *J. Thorac. Oncol.* **12**, 334–346 (2017).
6. Marchio, C. et al. Distinctive pathological and clinical features of lung carcinoids with high proliferation index. *Virchows Arch. : Int. J. Pathol.* **471**, 713–720 (2017).

7. Pelosi, G., Rindi, G., Travis, W. D. & Papotti, M. Ki-67 antigen in lung neuroendocrine tumors: unraveling a role in clinical practice. *J. Thorac. Oncol.* **9**, 273–284 (2014).
8. Derks, J. L. et al. New insights into the molecular characteristics of pulmonary carcinoids and large cell neuroendocrine carcinomas, and the impact on their clinical management. *J. Thorac. Oncol.* **13**, 752–766 (2018).
9. Pelosi, G. et al. Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. *Virchows Arch.* **472**, 567–577 (2018).
10. Rekhtman, N. et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma-like and non-small cell carcinoma-like subsets. *Clin. Cancer Res.* **22**, 3618–3629 (2016).
11. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
12. George, J. et al. Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nat. Commun.* **9**, 1048 (2018).
13. Peifer, M. et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **44**, 1104–1110 (2012).
14. George, J. et al. Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).
15. Argelaguet, R. et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
16. Straif, K. et al. A review of human carcinogens—Part C: metals, arsenic, dusts, and fibres. *Lancet Oncol.* **10**, 453–454 (2009).
17. Carbone, M. et al. BAP1 and cancer. *Nat. Rev. Cancer* **13**, 153–159 (2013).
18. Kiefer, J. et al. Abstract 3589: a systematic approach toward gene annotation of the hallmarks of cancer. *Cancer Res.* **77**, 3589–3589 (2017).
19. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
20. Shi, C. & Pamer, E. G. Monocyte recruitment during infection and inflammation. *Nat. Rev. Immunol.* **11**, 762–774 (2011).
21. Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. *Nat. Rev. Immunol.* **13**, 159–175 (2013).
22. Jakubzick, C. V., Randolph, G. J. & Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nat. Rev. Immunol.* **17**, 349–362 (2017).
23. Cernadas, M., Lu, J., Watts, G. & Brenner, M. B. CD1a expression defines an interleukin-12 producing population of human dendritic cells. *Clin. Exp. Immunol.* **155**, 523–533 (2009).
24. Odom, D. T. et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
25. Tran Janco, J. M., Lamichhane, P., Karyampudi, L. & Knutson, K. L. Tumor-infiltrating dendritic cells in cancer pathogenesis. *J. Immunol.* **194**, 2985–2991 (2015).
26. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: what we know, what we need to know and the path forward. *Nat. Rev. Cancer* **17**, 765 (2017).
27. Rudin, C. M. et al. Rovalpituzumab tesirine, a DLL3-targeted antibody-drug conjugate, in recurrent small-cell lung cancer: a first-in-human, first-in-class, open-label, phase 1 study. *Lancet Oncol.* **18**, 42–51 (2017).
28. Gara, R. K. et al. Slit/Robo pathway: a promising therapeutic target for cancer. *Drug Discov. Today* **20**, 156–164 (2015).
29. Boers, J. E., den Brok, J. L., Koudstaal, J., Arends, J. W. & Thunnissen, F. B. Number and proliferation of neuroendocrine cells in normal human airway epithelium. *Am. J. Respir. Crit. Care Med.* **154**, 758–763 (1996).
30. Sutherland, K. D. & Berns, A. Cell of origin of lung cancer. *Mol. Oncol.* **4**, 397–403 (2010).
31. Branchfield, K. et al. Pulmonary neuroendocrine cells function as airway sensors to control lung immune response. *Science* **351**, 707–710 (2016).
32. Kimura, H. et al. Randomized controlled phase III trial of adjuvant chemo-immunotherapy with activated killer T cells and dendritic cells in patients with resected primary lung cancer. *Cancer Immunol. Immunother.: CII* **64**, 51–59 (2015).
33. Scarpa, A. et al. Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature* **543**, 65–71 (2017).
34. Simbolo, M. et al. Lung neuroendocrine tumours: deep sequencing of the four World Health Organization histotypes reveals chromatin-remodelling genes as major players and a prognostic role for TERT, RB1, MEN1 and KMT2D. *J. Pathol.* **241**, 488–500 (2017).
35. Swarts, D. R. et al. CD44 and OTP are strong prognostic markers for pulmonary carcinoids. *Clin. Cancer Res.* **19**, 2197–2207 (2013).
36. Papaxoinis, G. et al. Prognostic significance of CD44 and orthopedia homeobox protein (OTP) expression in pulmonary carcinoid tumours. *Endocr. Pathol.* **28**, 60–70 (2017).
37. Koyama, T. et al. ANGPTL3 is a novel biomarker as it activates ERK/MAPK pathway in oral cancer. *Cancer Med.* **4**, 759–769 (2015).
38. Kurppa, K. J., Denessiouk, K., Johnson, M. S. & Elenius, K. Activating ERBB4 mutations in non-small cell lung cancer. *Oncogene* **35**, 1283–1291 (2016).
39. Williams, C. S. et al. ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. *Carcinogenesis* **36**, 710–718 (2015).
40. Fabbri, A. et al. Thymus neuroendocrine tumors with CTNNB1 gene mutations, disarrayed ss-catenin expression, and dual intra-tumor Ki-67 labeling index compartmentalization challenge the concept of secondary high-grade neuroendocrine tumor: a paradigm shift. *Virchows Arch.* **471**, 31–47 (2017).
41. Wang, T. T. et al. Tumour-activated neutrophils in gastric cancer foster immune suppression and disease progression through GM-CSF-PD-L1 pathway. *Gut* **66**, 1900–1911 (2017).
42. Mojic, M., Takeda, K. & Hayakawa, Y. The dark side of IFN-gamma: its role in promoting cancer immunoevasion. *Int. J. Mol. Sci.* **19**, pii: E89 (2017).
43. Zaidi, M. R. & Merlino, G. The two faces of interferon-gamma in cancer. *Clin. Cancer Res.* **17**, 6118–6124 (2011).
44. Ocana, A., Nieto-Jimenez, C., Pandiella, A. & Templeton, A. J. Neutrophils in cancer: prognostic role and therapeutic strategies. *Mol. cancer* **16**, 137 (2017).
45. Tang, L. H., Basturk, O., Sue, J. J. & Klimstra, D. S. A practical approach to the classification of WHO grade 3 (G3) well-differentiated neuroendocrine tumor (WD-NET) and poorly differentiated neuroendocrine carcinoma (PD-NEC) of the pancreas. *Am. J. Surg. Pathol.* **40**, 1192–1202 (2016).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
47. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
50. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
51. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* **30**, 2813–2815 (2014).
52. Delhomme, T. M. et al. needlestack: an ultra-sensitive variant caller for multi-sample deep next generation sequencing data. *bioRxiv* https://doi.org/10.1101/639377 (2019).
53. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.* **38**, e164 (2010).
54. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr. Proto. Bioinforma.* **51**, 11.14.11-19, https://doi.org/10.1002/0471250953.bi1114s51 (2015).
56. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
57. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
58. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
59. Fernandez-Cuesta, L. et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* **16**, 7 (2015).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
61. Dray, S. & Dufour, A. B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 20 (2007).
62. Finotello, F. et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
63. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
64. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucl. Acids Res.* **44**, e20 (2016).
65. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* **43**, e47 (2015).
66. Mo, Q. et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl Acad. Sci. USA* **110**, 4245–4250 (2013).
67. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).

68. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
69. Zhang, J. D. & Wiemann, S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470–1471 (2009).
70. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

Acknowledgements

We thank the patients donating their tumour specimens. We also thank Prof. Roman K. Thomas, Dr. Martin Peifer, Dr. Julie George, Dr. Paul Brennan, and Dr. Ghislaine Scelo for their help with logistics. We also thank Dr. Ricard Argelaguet for his advice in using MOFA. This study is part of the lungNENomics project and the Rare Cancers Genomics initiative (www.rarecancersgenomics.com). This work has been funded by the US National Institutes of Health (NIH R03CA195253 to L.F.C. and J.D.M.), the French National Cancer Institute (INCa, PRT-K-17-047 to L.F.C. and TABAC 17-022 to J.D.M.), the Ligue Nationale contre le Cancer (LNCC 2016 to L.F.C.), France Genomique (to J.D.M.), and the Italian Association for Cancer Research (AIRC) (IG 19238 to M.V. and MFAG 12983 to L.A.M.) (Special Programme 5X1000, ED No12162 to U.P., L.R., and G.S.). J.S. is a Miguel Servet researcher (CP13/00055 and PI16/0295). L.M. and T.M.D. have fellowships from the LNCC.

Author contributions

L.F.C. conceived and designed the study. L.F.C. and M.F. supervised all the aspects of the study. A.G., A.B., J.A., F.L.C.K., S.B., J.S., N.G. and S.Lan. supervised some aspects of the study. B.A.A., E.B. and S.Lan. performed the histopathological review. N.Leb., T.G., J.D., A.C., C. Cu., G.D. and N.Lem. did the lab work. N.A., N.Leb., A.A.G.G., L.M., D.H., A.S.S., A.F., T.M.D., R.O., V.M., C.V. and L.A.M. performed the computational and statistical analyses. P.L., A.C.T., A.S., J.H.C., J. Saenger, J. Stojsic, J.K.F., M.B., C.B.F., F.G.S., N.L.S., P.A.R., G.W., L.R., G.S., U.P., M.M., S.Lac., J.M.V., V.H., P.H., O.T.B., M.L.-I., V.T.M., L.A.M., P.G., M.V., M.G.P., L.B., H.P., A.M.C.D., E.B., E.J.M.S., N.G. and S.Lan contributed with samples and the corresponding histopathological, epidemiological, and clinical data. J.F.D., Z.H., A.V., P.N. and J.D.M. helped with logistics. J.D., B.A.A., C. Ca., L.R., M.M., M.V., M.G.P., L.B., H.P., G.P., J.D.M., H.H.V., E.J.M.S., N.G. and S.Lan gave scientific input. N.A., N.Leb., A.A.G.G., L.M., J.D.M., M.F. and L.F.C. wrote the manuscript, which was reviewed and commented by all the co-authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11276-9.

Competing interests: The authors declare no competing interests. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organisation, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organisation.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: *Nature Communications* thanks Florian Buettner and Takashi Kohno for their contribution to the peer review of this work. Peer reviewer reports are available

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

N. Alcala[1,35], N. Leblay[1,35], A.A.G. Gabriel[1,35], L. Mangiante[1], D. Hervas[2], T. Giffon[1], A.S. Sertier[3], A. Ferrari[3], J. Derks[4], A. Ghantous[5], T.M. Delhomme[1], A. Chabrier[1], C. Cuenin[5], B. Abedi-Ardekani[1], A. Boland[6], R. Olaso[6], V. Meyer[6], J. Altmuller[7], F. Le Calvez-Kelm[1], G. Durand[1], C. Voegele[1], S. Boyault[8], L. Moonen[4], N. Lemaitre[9], P. Lorimier[9], A.C. Toffart[10], A. Soltermann[11], J.H. Clement[12], J. Saenger[13], J.K. Field[14], M. Brevet[15], C. Blanc-Fournier[16], F. Galateau-Salle[17], N. Le Stang[17], P.A. Russell[18], G. Wright[18], G. Sozzi[19], U. Pastorino[19], S. Lacomme[20], J.M. Vignaud[20], V. Hofman[21], P. Hofman[21], O.T. Brustugun[22,23], M. Lund-Iversen[23], V. Thomas de Montpreville[24], L.A. Muscarella[25], P. Graziano[25], H. Popper[26], J. Stojsic[27], J.F. Deleuze[6], Z. Herceg[5], A. Viari[3], P. Nuernberg[7,28], G. Pelosi[29], A.M.C. Dingemans[4], M. Milione[19], L. Roz[19], L. Brcic[26], M. Volante[30], M.G. Papotti[30], C. Caux[31], J. Sandoval[2], H. Hernandez-Vargas[32], E. Brambilla[9], E.J.M. Speel[4], N. Girard[33,34], S. Lantuejoul[3,8,17], J.D. McKay[1], M. Foll[1] & L. Fernandez-Cuesta[1]

[1]International Agency for Research on Cancer (IARC/WHO), Section of Genetics, 150 Cours Albert Thomas, 69008 Lyon, France. [2]Health Research Institute La Fe, Avenida Fernando Abril Martorell, Torre 106 A 7planta, 46026 Valencia, Spain. [3]Synergie Lyon Cancer, Centre Léon Bérard, 28 Rue Laennec, 69008 Lyon, France. [4]Maastricht University Medical Centre (MUMC), GROW School for Oncology and Developmental Biology, P.O. Box 5800, 6202 AZMaastricht, The Netherlands. [5]International Agency for Research on Cancer (IARC/WHO), Section of Mechanisms of Carcinogenesis, 150 Cours Albert Thomas, 69008 Lyon, France. [6]Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, 2 rue Gaston Crémieux, CP 5706, 91057 Evry Cedex, France. [7]Cologne Centre for Genomics (CCG) and Centre for Molecular Medicine Cologne (CMMC), University of Cologne, Weyertal 115, 50931 Cologne, Germany. [8]Translational Research and Innovation Department, Cancer Genomic Platform, 28 Rue Laennec, 69008 Lyon, France. [9]Institute for Advanced Biosciences, Site Santé, Allée des Alpes, 38700 La Tronche, Grenoble, France. [10]Pulmonology—Physiology Unit, Grenoble Alpes University Hospital, 38700 La Tronche, France. [11]Institute of Pathology and Molecular Pathology, University Hospital Zurich, Schmelzbergstrasse 12, 8091 Zurich, Switzerland. [12]Department Hematology and Medical Oncology, Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany. [13]Bad Berka Institute of Pathology, Robert-Koch-Allee 9, 99438 Bad Berka, Germany. [14]Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, 6 West Derby Street, L7 8TX Liverpool, UK. [15]Pathology Institute, Hospices Civils de Lyon, University Claude Bernard Lyon 1, 59 Boulevard Pinel, 69677 BRON Cedex, France. [16]CLCC François Baclesse, 3 avenue du Général Harris, 14076

Caen Cedex 5, France. [17]Department of Pathology, Centre Léon Bérard, 28, rue Laennec, 69373 Lyon Cedex 8, France. [18]St. Vincent's Hospital and University of Melbourne, Victoria Parade, Fitzroy, Melbourne, VIC 3065, Australia. [19]Pathology Division Fondazione, IRCCS Istituto Nazionale dei Tumori, Via G. Venezian 1, 20133 Milan, Italy. [20]Nancy Regional University Hospital, CHRU, CRB BB-0033-00035, INSERM U1256, 29 Avenue du Maréchal de Lattre de Tassigny, 54035 Nancy Cedex, France. [21]Laboratory of Clinical and Experimental Pathology, FHU OncoAge, Nice Hospital, Biobank BB-0033-00025, IRCAN Inserm U1081 CNRS 7284, University Côte d'Azur, 30 avenue de la voie Romaine, CS, 51069-06001 Nice Cedex 1, France. [22]Drammen Hospital, Vestre Viken Health Trust, Vestre Viken HF, Postboks 800, 3004 Drammen, Norway. [23]Institute of Cancer Research, Oslo University Hospital, Ullernchausseen 70, 0379 Oslo, Norway. [24]Marie Lannelongue Hospital, 133 avenue de la Resistance, 92350 Le Plessis Robinson, France. [25]Fondazione IRCCS Casa Sollievo della Sofferenza, Viale Cappuccini 1, 71013 San Giovanni Rotondo FG, Italy. [26]Diagnostic and Research Institute of Pathology, Medical University of Graz, Neue Stiftingtalstrasse 6, 8010 Graz, Austria. [27]Department of Thoracopulmonary Pathology, Service of Pathology, Clinical Center of Serbia, Pasterova 2, Belgrade 11000, Serbia. [28]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Joseph-Stelzmann-Straße 26, 50931 Cologne, Germany. [29]Department of Oncology and Hemato-Oncology, University of Milan, and Inter-Hospital Pathology Division, IRCCS Multimedica, Via Gaudenzio Fantoli, 16/15, 20138 Milan, Italy. [30]Department of Oncology, University of Turin, Pathology Division, Via Santena 7, 10126 Torino, Italy. [31]Department of Immunity, Virus, and Inflammation, Cancer Research Centre of Lyon (CRCL), 28 Rue Laennec, 69008 Lyon, France. [32]Cancer Research Centre of Lyon (CRCL), Inserm U 1052, CNRS UMR 5286, Centre Léon Bérard, Université de Lyon, 28 Rue Laennec, 69008 Lyon, France. [33]Institut Curie, 26 Rue d'Ulm, 75005 Paris, France. [34]European Reference Network (ERN-EURACAN), 28 rue Laennec, 69008 Lyon, France. [35]These authors contributed equally: Alcala N., Leblay N., Gabriel A. A. G. [36]These authors jointly supervised this work: Foll M., Fernandez-Cuesta L.

Chapter 3

Generation of a pan-LNEN tumor map using data integration

3.1 Context

The rise of large genomics studies have not only continually led to a multitude of molecular discoveries but have also increased the reuse potential of molecular datasets. In this context, computational tools have been used to perform data integration. Such studies attempt to increase sample sizes as well as to contrast the molecular profiles of tumors to provide new insights with potential applications in the clinic. Firstly, the growth of initiatives like the TCGA and ICGC initiatives have enabled researchers to access genomics data from different tumor types and to move towards cross-cancer studies like the Pan-Cancer Atlas project [51]. These studies have the advantage to increase statistical power, for example, to identify cancer genes mutated at intermediate frequencies [133], and allow to explore similarities and contrasts between tumor types [75]. Tools recently used in pan-cancer studies are molecular maps. They result from the integration of large datasets containing thousands of molecular variables, called features (*e.g.* expression or methylation levels, mutations or copy number variations) that have been embedded in a lower dimensional representation of the molecular data. The integration of such datasets, when coming from diverse and heterogeneous studies, require harmonized data processing to enable the comparison of samples. As such molecular maps can be considered as visualization and interpretation tools that are based on a complex but homogenized set of features. Using such tools and data, diverse oncogenic processes in tumors from distinct tissues have been observed [122, 134]. Hoadley *et al.* also identified samples whose initial classification, mainly based on the tumor tissue of origin, was in contradiction with their molecular taxonomy, hence revealing potential misclassifications [122]. It has been estimated that one out of ten samples could be reclassified based on their molecular profile [122]. In addition to highlighting the

molecular diversity of cancers, molecular maps have also been used to reveal similarities between tumors with, for example, the identification of groups constituted of a mixture of tissue subtypes by Bolouri *et al.* [135], a mixture of cancers from similar cell of origin by Hoadley *et al.* [134] and the identification of a pan-cancer group by Newton *et al.* [136].

The integration of omics datasets to produce such maps brings out however multiple challenges. Firstly, as mentioned in the Findable, Accessible, Interoperable, Reusable (FAIR) principles [137], the data should be accessible and reusable. Secondly, the data need to be comparable. When the data to integrate come from multiple studies, samples are often sequenced in distinct centers and processed using different protocols. In the case of large consortia like the TCGA, large efforts of homogenization have been made [62]. For smaller studies, available on different data repositories, data harmonization is still required before integration to limit batch effects. A common preprocessing workflow, using the same methods, the same software versions, the same machines have to be developed.

As mentioned in the previous chapter (Chapter 2), in the case of rare cancers like the lung neuroendocrine tumors, the sample size of molecular studies is limited. Hence, a higher number of studies or the design of larger studies are required. Integrating the datasets already published is a first step in that direction. In the work presented in this chapter, we generated a molecular map of the LNEN tumors by integrating datasets from six studies and provided multiple resources to reproduce and expand the molecular map in the future.

3.2 Research contribution

3.2.1 Introduction

The previous chapter (Chapter 2) presented the lung neuroendocrine tumor types as distinct diseases in terms of etiology, clinical characteristics but also in terms of molecular profiles. The use of multi-omics data identified new molecular groups of pulmonary carcinoids that were not perfectly matching the histopathological classification. Those molecular groups were clinically relevant as they had different prognosis. The pulmonary carcinoids were stratified in three clusters A1, A2 and B and the supra-carcinoids, a subgroup of carcinoids clustering with the LCNEC samples, were unveiled. However, the sample size of these groups is limited, especially the supra-carcinoids that formed a group of six samples. The molecular diversity of the LNEN samples hence still needs to be further explored. Increasing the sample size

of future studies would, on one hand, enable the validation of the newly identified entities and on the other hand, the identification of new subgroups. Considering the rarity of LNEN cancers, those objectives could be reached by integrating available datasets from as wide a range of sources as possible, including for example public data archives.

Also, the identification of the supra-carcinoids, as well as carcinoids samples with intermediate profiles (Chapter 2), suggested that the lung neuroendocrine subtypes could share more links than expected, which is a hypothesis supported by previous molecular studies [95, 138]. Hence, integrating datasets from the different LNEN subtypes could provide further evidence of those links and visualizing each sample in the context of other tumor types could lead to new hypotheses.

In order to visualize omics datasets, which interpretation is complex due to the large number of features (thousands of features, *e.g.* genes expression or methylation levels), adapted computational methods, like dimensionality reduction methods are needed (See section 1.4). A common method used to visualize data in lower dimensions is PCA. However, PCA decomposes the data in multiple principal components, each explaining a certain part of the variance in the data. To represent the initial data in two dimensions for human visualization and interpretation, only two principal components need to be chosen. The signals captured by the other axes are thus lost. Also, the PCA method does not capture non-linear structure in the data. In this work, we chose to use another dimensionality reduction method that overcomes these issues, the Uniform Manifold Approximation and Projection (UMAP) presented in Section 1.4, to generate a two-dimensional molecular map of the LNEN tumors.

In this work, we took advantage of the study presented in the previous chapter by reusing the transcriptomic data generated and integrating them with five other datasets to build a molecular map of the LNEN samples. The preprocessing and quality control steps performed on the first dataset were described and reused for data homogenization, and resources to promote and facilitate further use of the data were provided.

3.2.2 Material and methods

Supplementary description of the LNEN data presented in Chapter 2

The study presented in the previous chapter 2 generated omics data of lung NEN samples [139]. In total, whole exome and whole genome data were generated for 16 and 3 samples respectively, transcriptomic data for 20 samples and EPIC 850k arrays

data for 76 samples. The first part of the paper presented in this chapter provided a complementary description of these data. The aim was to facilitate the reuse of the data and promote its integration with other datasets. Data from different studies are generated under different conditions (different machines, protocols). Before any integration analysis, it is thus necessary to preprocess each data set in the same way and to assure that their quality is homogeneous to avoid batch effects in subsequent analyses. Hence, common quality controls (QC) and preprocessing protocols are required. We described the preprocessing steps performed on the previously mentioned data as well as their quality: i) sequence qualities for WGS, WES and RNA sequencing, ii) quality of the DNA and RNA sequencing reads alignment, iii) quality of the RNA sequencing reads assignments to genes, and iv) the quality of the methylation arrays.

Integration of additional datasets

In a second part, the transcriptomic data from the previous dataset (See [139] and Chapter 2), were integrated with other transcriptomic data generated by previous studies characterizing the molecular patterns of different types of lung NEN tumors. Pulmonary carcinoids (mostly typical carcinoids) have been described in 2014 by Fernandez-Cuesta *et al.* [92] and in 2019 by Laddha *et al.* [140]. In 2018, the expression patterns of LCNEC samples were described by Georges *et al.* [93]. The genome and the transcriptome of SCLC samples have been explored by Rudin *et al.* [128] and George *et al.* [87]. For each dataset, transcriptomic data were available on the EGA [128, 92, 87, 93, 139] or GEO [140] data repositories. In total, six transcriptomic datasets were gathered and for the purpose of homogenization, processed in the same way following the steps described in the first part of the paper. The pipelines used were coded using the workflow management system Nextflow [141] and can be run using containerization tools like Docker [142] and Singularity [143]. Nextflow allowed us to organize the several processing steps in a completely automatized and reproducible pipeline. Containers are virtual machines that allow to embed the required computing environment with all the needed softwares. It thus firstly assures that the analyses are reproducible since the same softwares and versions will be used to reprocess the data. Secondly, it provides portability, *i.e.* the pipeline will run similarly on heterogeneous computing environments. The workflows, whose development was made in house and led by Dr. Alcala, are hosted on GitHub [144]. The combination of all these tools enables future users to run the

pipelines using simplified command lines. Figure 3.1 provides the nextflow commands needed to process RNA-Seq data from reads mapping to gene expression quantification, following the steps used in this study.

① **RNA sequencing mapping and quality controls**

```
nextflow run iarcbioinfo/RNAseq-nf -r v2.3 -profile singularity
--input_folder folder_with_fastq_files --output_folder out_RNAseq-nf-2.3
--ref_folder ref_genome.fa.star.idx/ --ref ref_genome.fa --gtf
ref_annot.gtf --bed hg38_Gencode_v33.bed --fastq_ext fastq
--STAR_mapqUnique 60 --cutadapt
```

② **BAM realignment**

```
nextflow run iarcbioinfo/abra-nf -r v3.0 -profile singularity --bam_folder
out_RNAseq-nf-2.3/BAM --output_folder out_abra-nf-3.0 --ref ref_genome.fa
--gtf ref_annot.gtf --bed hg38_Gencode_v33_merged.bed --junctions --rna
```

③ **Base quality score recalibration**

```
nextflow run iarcbioinfo/BQSR-nf -r v1.1 -profile singularity
--input_folder out_abra-nf-3.0/ --ref ref_genome.fa --snp_vcf
dbsnp_146.hg38.vcf.gz --indel_vcf
Mills_and_1000G_gold_standard.indels.hg38.vcf.gz --output_folder out_bqsr-
nf-1.1
```

④ **Gene expression quantification**

```
nextflow run iarcbioinfo/RNAseq-transcript-nf -r v2.1 -profile singularity
--input_file input-transcript.txt  --gtf ref_annot.gtf --output_folder
out_RNAseq-transcript-nf-2.1
```

FIGURE 3.1: **Nextflow command lines to perform RNA-Seq preprocessing.** Where: *folder_with_fastq_files* is a folder with the fastq files to process; *ref_genome.fa.star.idx* is a folder with the genome reference files for the software STAR; *ref_annot.gtf* the annotation file; *hg38_Gencode_v33.bed* a bed file with a list of intervals for further annotations by RSeQC; *dbsnp_146.hg38.vcf.gz* and *Mills_and_1000G_gold_standard.indels.hg38.vcf.gz* are Variant Call Format (VCF) files coming from the Genome Analysis Toolkit (GATK) bundle hg38; *input-transcript.txt* is a file containing the samples ID, paths to Binary Alignment Map (BAM) files, and read lengths per sample.

After homogenization of the six transcriptomic datasets, we built a pan-LNEN molecular map using the Uniform Manifold Approximation and Projection (UMAP) method [111]. UMAP is a dimensionality reduction method that is adapted, unlike PCA, to capture non-linear dependencies in the data. The algorithm is based on topology theory and follows two main steps. Firstly, it builds a graphical representation of the high dimensional data and secondly finds a simpler representation of the same graph in a lower space (See section 1.4 for more details). Reducing a

dataset composed of more than 50,000 dimensions to two dimensions systematically distorts the initial data structure and hinders the preservation of the original distances between samples. UMAP optimizes the dimensionality reduction to retain as much as possible the main structures, local or global structures, depending on the parameters chosen. According to the UMAP documentation [145], the most important parameter that will determine if global or local structure will be preserved is the *n_neighbors* parameter, which is the number of nearest neighbors considered in the model to build the high-dimensional graph (See section 1.4). The higher the parameter value, the higher is the number of connected points in the initial graph, hence the better is the preservation of the global structure (See Figure 3.2). Considering this parameter's influence, we compared the UMAP representations obtained using respectively the *n_neighbors* parameter default value of 15 and fixing the *n_neighbors* parameter to 238, which is the total number of samples. This comparison showed that the latter parameter choice led to better preservation of the global structure of the data and was thus chosen for the final molecular map.

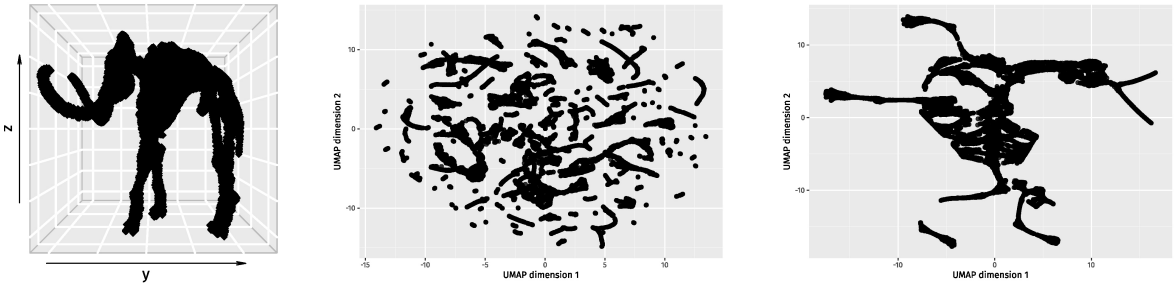


FIGURE 3.2: **Illustration of the influence of the *n_neighbors* parameter on UMAP representations.** Representations of a mammoth skeleton using UMAP (Data retrieved from the following GitHub repository: MNoichl/UMAP-examples-mammoth-). Left panel: 3D representation of the original dataset. Middle and right panels: UMAP two dimensional representations fixing the parameter *n_neighbors* to 15 (UMAP default value) and 100 respectively.

After producing the pan-LNEN molecular map, its quality was assessed using three analyses. Firstly, we verified that the six molecular clusters previously established in the different studies were re-identified and tested if biological hypotheses generated by the same studies could be reformulated using the molecular map. More specifically, in the different studies, samples have been reported to have discordant molecular and histopathological profiles. For those samples, we computed and compared, on the pan-LNEN molecular map, their Euclidean distances to the centroid of their molecular cluster and the centroid of their histopathological group.

We then evaluated whether the samples neighbourhoods in the original high dimensional space were preserved in low dimensional representation generated by UMAP. For that purpose, we used the sequence difference (SD) metric, defined by Martins *et al.* [146] as follows:

$$SD_k(i) = \frac{1}{2} \sum_{j \in V_k^1(i)} [k - \rho_i^1(j)].|\rho_i^1(j) - \rho_i^2(j)| + \frac{1}{2} \sum_{j \in V_k^2(i)} [k - \rho_i^2(j)].|\rho_i^1(j) - \rho_i^2(j)| \tag{3.1}$$

This metric compares, for each sample i, the k nearest (using Euclidean distances) samples in the neighborhoods, $V_k^1(i)$ and $V_k^2(i)$, of two spaces D_1 and D_2 respectively. The metric evaluates the rank of each neighbor j in the two spaces ($\rho_i^1(j)$ and $\rho_i^2(j)$) and penalizes discordant rankings, a higher penalty being added for close neighbors. Hence, values of SD close to zero indicates the preservation of the samples neighborhood. To assess the quality of the UMAP projection, we compared UMAP to PCA dimensionality reduction results that we considered as references. The SD metric was thus used to compare the samples neighborhoods in the original space and UMAP representations as well as in PCA representations based on the two and five first principal components (PC) respectively (PCA-2D and PCA-5D). We expect PCA-2D to perform poorly in contrast to PCA-5D since the best projection to capture the six molecular groups, previously identified in the different studies, would have five dimensions. For the UMAP method, two representations were considered, one fixing the *n_neighbors* parameters to the default value 15 and the other to 238. For each comparison, we computed the mean SD values across all samples while varying the parameter k to assess samples preservation at various scales, from local to global scale. Note that this analysis guided the choice of the *n_neighbors* parameter described previously.

Finally, we evaluated whether the molecular map was able to retain the gene expression structure in the original space. For that sake, the Moran index (MI), which is a spatial-autocorrelation measure, was used. Genes with expression varying randomly across the map will have an MI value of 0, genes with dissimilar expression levels in nearby regions a value of -1 (negative auto-correlation) and genes with similar expression levels in close samples a value of 1 (positive autocorrelation). The Moran index value of each gene in the original space, the PCA-5D and the UMAP projections were computed, and the top-ranking genes in the three spaces compared.

3.2.3 Results

Supplementary description of the LNEN data presented in Chapter 2

In the first part of the paper, the pipelines used to preprocess the data, as well as the associated quality controls, were described. More specifically, for the RNA-Seq data from Alcala *et al.* included in the molecular map, the reads bases and reads assignments quality standards have been satisfied. All samples had: i) high means per base sequence qualities (above 28), ii) more than 70% of reads were uniquely mapped, iii) more than 75% of reads mapped in coding regions, and iv) more than 70% of reads were assigned to genes from the reference annotations.

A molecular map based on transcriptomic datasets integration

In this study, six RNA-Seq datasets were available and were processed homogeneously for data integration. Based on the harmonized dataset, a two-dimensional UMAP representation of the pan-LNEN samples was obtained and is represented in Figure 3.3. The map revealed distinct clusters of samples matching the molecular clusters previously identified in the respective studies. Firstly, out of the six studies, two from Alcala *et al.* [139] (See Chapter 2) and from Laddha *et al.* [140], identified three clusters of carcinoids samples. Those three carcinoids clusters were also distinguishable on the molecular map. Moreover, the clusters of the two studies matched perfectly: the clusters A1, A2 and B from the first study correspond respectively to the clusters LC1, LC3 and LC2 from the latter (See Figure 3.3). Also, the LCNEC samples were split into two groups, the previously identified LCNEC type-I and LCNEC type-II samples [93]. Finally, the supra-carcinoids, which have the morphology of carcinoids but the molecular features of LCNEC [139] as well as one sample from the LC2 group, clustered with the LCNEC samples.

In addition, previous biological hypotheses were consistent on the pan-LNEN map. For example, in 2018, George *et al.* [93] defined the SCLC-LCNEC like samples as histological SCLC samples having an LCNEC molecular profile and the LCNEC-SCLC like samples as histological LCNEC samples having an SCLC molecular profile. On the pan-LNEN map the SCLC-LCNEC like samples were closer to the LCNECs than the SCLCs and the LCNEC-SCLC like samples clustered with the SCLCs rather than with the LCNEC samples.

To further assess the quality of the molecular map, the preservation of samples neighborhood and spatial auto-correlations by UMAP were evaluated. To determine if the samples neighborhoods were correctly preserved, we computed for each

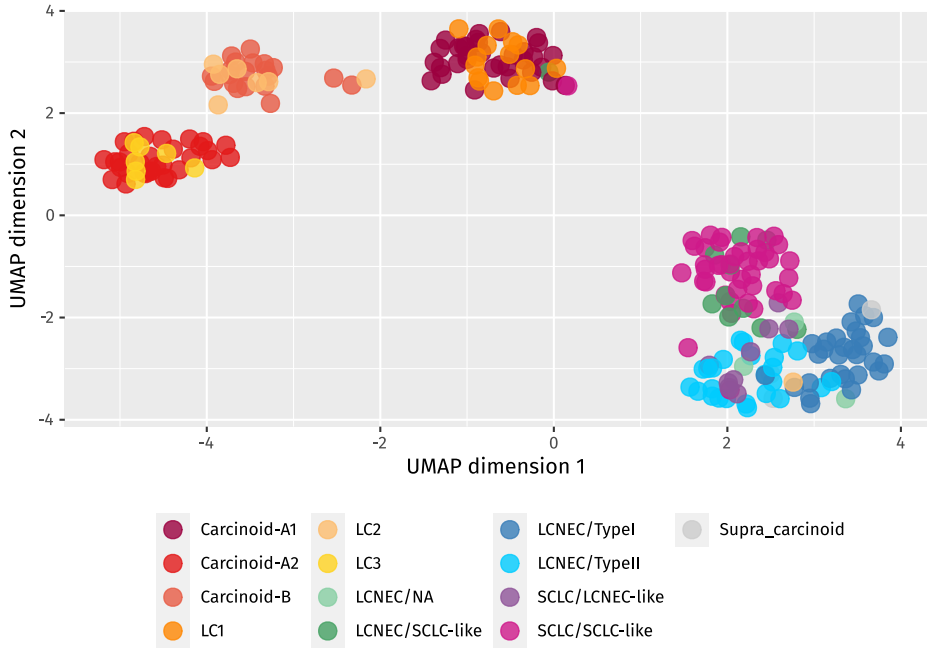


FIGURE 3.3: **The pan-LNEN molecular map**. The x and y-axis represent respectively the first and second dimensions resulting from UMAP dimensionality reduction. Each dot corresponds to one sample; the colors describe the molecular clusters identified by the previous studies.

sample the SD metric (Equation 3.1) that allows to measure the dissimilarity between the sample's neighborhood in the original space and the sample's neighborhood in lower dimensional representations. UMAP provided a trade-off between the ability to visualize the samples in two dimensions and the conservation of the structure in the data since the method improved the samples neighborhood preservation over the PCA-2D representation. When comparing the UMAP representations based on different *n_neighbors* parameters, 15 (UMAP-15) and 238 (UMAP-238), we observed that UMAP-15 had a better preservation when considering local neighborhoods only, while UMAP-238 outperformed for global preservation. Finally, based on the MI measure, the preservation of spatial auto-correlation was tested. The genes with the highest MI values in different representations of the data - the original space, the PCA-5D representation and the UMAP representation - were concordant. Indeed, 88.8% overlap between the three sets of top 1000 genes was observed.

Along with the paper, multiple resources have been produced to facilitate the reuse of the data and the integration of future datasets. The reproducible workflows used for the data preprocessing are available on GitHub [144]. The paper is accompanied by a computational notebook on Nextjournal [147] (See notebook here), which provides the nextflow command lines and the reference files used to perform

the RNA-Seq data integration as well as the code needed to reproduce the results presented in the paper. This integrative notebook could be reused to test different parameters and evaluate their influences on the results. Finally, the molecular map was uploaded on TumorMap [136] (See pan-LNEN tumor map here), which is a genomics portal based on Google Maps technology to enable genomics data visualization and exploration, and thus foster the generation of new hypotheses. External users can indeed interactively manipulate the molecular map, *e.g.* by selecting subsets of samples, manipulating metadata in order to identify variables exhibiting interesting distributions on the map or performing statistical tests. Figure 3.4 presents an example of the pan-LNEN map representation on TumorMap, when choosing to represent *MKI67* expression levels distribution across the samples.

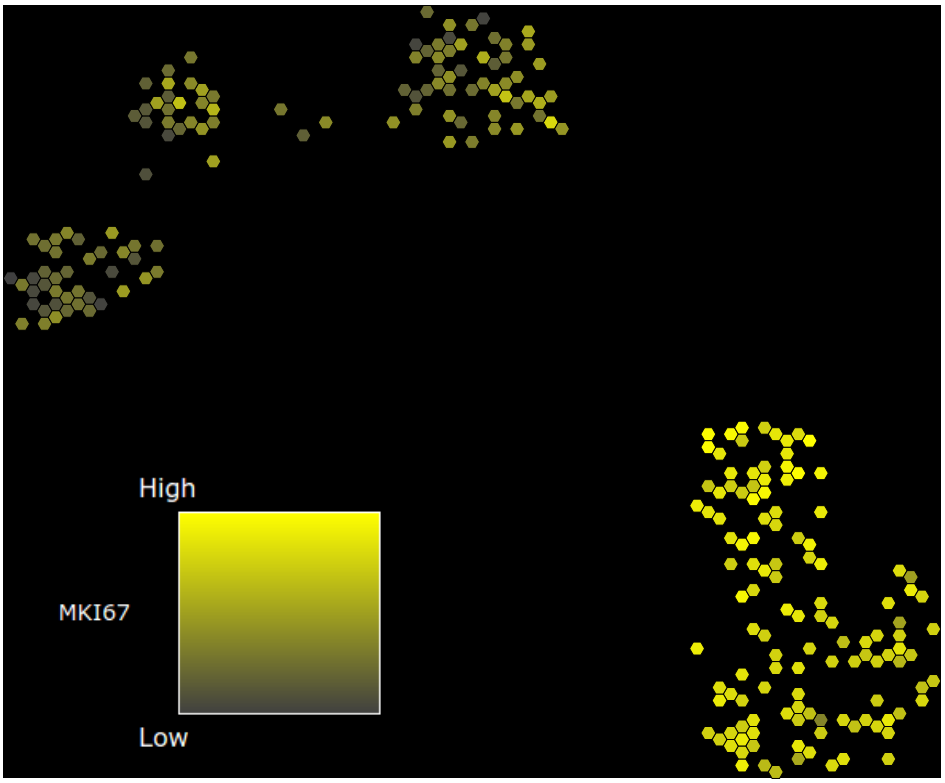


FIGURE 3.4: **Example of the pan-LNEN representation on TumorMap**. Each dot corresponds to one sample; the color gradient represents the expression levels of *MKI67*. Note that the samples with the highest *MKI67* expression levels (bottom right corner) are the SCLCs and LC-NECs.

3.2.4 Conclusion and discussion

In this study, we described the quality of the genomics data published by Alcala *et al.* in 2019 and the preprocessing steps to follow to enable their integration with other

datasets. Using the described workflows, we integrated the transcriptomic data from this previous study with five other datasets to build the first two-dimensional pan-LNEN molecular tumor map. Its quality was assessed by evaluating the preservation of samples neighborhood and spatial auto-correlations, and by validating previously identified molecular clusters. Finally, we provided the homogenized data underlying the map as well as distinct resources to promote further integration and exploration of these datasets.

The pan-LNEN molecular map presented in this chapter was based on the integration of six transcriptomic datasets that led to the identification of distinct LNEN molecular groups in previous studies. Two of them identified three groups of carcinoids [139, 140]. On the pan-LNEN molecular map, the carcinoids groups found in the two independent studies were consistent, suggesting that residual inter-study variations are not the major sources of variations captured by the map. Note that running PCA on the harmonized data supported this observation (See Figure in Annex B). The two studies might thus have identified the same carcinoids entities. This result reflects how data integration could be valuable for research purposes not only to reveal new molecular groups and to generate new hypotheses but also to confirm previous discoveries, especially clustering results whose biological relevance are not always easy to validate [99].

Besides their research relevance, molecular maps could be helpful in a clinical setting. Indeed, some samples can be difficult to classify in one cancer type category, or the tissue of origin of a tumor can be unknown, which complicates the diagnosis and subsequent treatment options for the patients. In those cases, we could imagine using the molecular map as a reference, project unknown samples on the map and determine with which molecular group, the samples best fit [136]. This option faces some limitations though. Firstly, the projection of samples would require to have a fixed map. However, in our case, the molecular diversity of the LNEN might not yet be fully discovered and larger studies would probably lead to a different structure. As such, the reference map must be sufficiently robust to allow subsequent projection of a given patient sample. Also, UMAP method is based on a stochastic optimization step and is sensitive to parameter choices which can be an issue if a fixed and stable map is required. Secondly, dimensionality reduction methods are sensitive to batch effects that could remain even after a homogenized data preprocessing. The samples preservation method is an example highlighting this problem. Often, samples are stored using the Formalin-Fixed Paraffin-Embedded (FFPE) method. The method is known to degrade DNA and RNA and resulting data are difficult to compare with frozen tissue data. Since our map is based on the latter, further

development would be needed to integrate them with FFPE samples. Also, in our map, only RNA-Seq datasets were integrated. However, other omics data could be used to generate the map. In the context of methylation datasets, which are known to be prone to stronger batch effects, additional steps might be needed to correct for those effects. Indeed, the correlation between batch effect variables (*e.g* center, plate, array position) and the factors resulting from the dimensionality reduction used to produce the map could be tested prior to downstream analyses. Also, non-negative control probes can be used to generate surrogate variables, which should capture most of the variability associated to the control probes and could be used as covariates in downstream statistical analyses.

We showed that the UMAP representation managed to provide a two-dimensional representation of the LNEN data while preserving its global structure. However, this trade-off was highly influenced by the parameters chosen and the interpretation of the molecular map should be performed with caution. For example, even if UMAP can be tuned to retain as much as possible the global structure, the method, like the t-SNE method, uses local distances and distort the initial space to build the low dimensional representation. Hence, distances between clusters and the cluster's spread could be misleading [148]. New methods, den-SNE and densMAP, that overcome the latter issue have been recently developed [149] and might replace the initial t-SNE and UMAP methods. Also, one limitation of our methods to validate the quality of the molecular map is that the metrics applied used euclidean distances. Since the ability of the Euclidean distance to discriminate nearest from farther neighbors is weak in high dimensional data [99], the metrics might need to be adapted.

The integration of the six transcriptomic datasets to generate the pan-LNEN map required their download from two data repositories and the reprocessing of each dataset, which also requires good computational infrastructures and skills. The work presented in this chapter resulted in the sharing of a processed and homogenized LNEN dataset. While we provide reproducible pipelines that could be used to replicate the data processing and analyses, this dataset could be directly integrated with new datasets processed with the same pipelines, which saves both data storage and computation time. The pipelines are also portable and can thus be run on any computational environment. Yet, depending on the samples sizes, this step still requires access to computational resources that are not available in all research groups. One solution could be to use the cloud model [150]. With the increase of the genomics studies these last years, the size of the public data archives have largely

increased with a size doubling every 18 months [150]. Cloud computing could facilitate data reuse by avoiding duplication of datasets, easy pipeline reuse thanks to containerization tools (docker, singularity) and foster collaborations while ensuring respect of data privacy. Such alternatives have been already put in place for large studies like the TCGA. The GDC [61] hosting the data is indeed closely linked to cloud platforms like the CGC [151]. We can imagine that such environments will be further developed and used in the field of genomics in the future.

3.2.5 Contribution

In this work, I contributed to the data analyses which consisted in producing the pan-LNEN molecular map and evaluating the quality of this map. Another major aim of this paper was to promote the reuse of the data by the research community. I contributed to this goal by describing in the paper the processing steps and quality controls performed as well as by providing different resources to enable reproducibility and reuse. Among these resources, I contributed to the code and data provided on the IARCbioinfo/DRMetrics GitHub repository and to the computational notebook in Nextjournal. Finally, I had a major role in the redaction of the paper and its review.

3.3 Article 2: A molecular map of lung neuroendocrine neoplasms

GigaScience, 2020, 1–10

doi: xx.xxxx/xxxx
Manuscript in Preparation
Data Note

DATA NOTE

A molecular map of lung neuroendocrine neoplasms.

Aurélie AG Gabriel[1,†], Emilie Mathian[1,†], Lise Mangiante[1], Catherine Voegele[1], Vincent Cahais[2], Akram Ghantous[2], James D McKay[1], Nicolas Alcala[1], Lynnette Fernandez-Cuesta[1,†] and Matthieu Foll[1,*,†]

[1]Section of Genetics, International Agency for Research on Cancer (IARC-WHO), Lyon, France and [2]Section of Mechanisms of Carcinogenesis, International Agency for Research on Cancer (IARC-WHO), Lyon, France

[*]**Correspondence address.** Matthieu Foll, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France. E-mail: follm@iarc.fr
ORCID IDs: Aurélie AG Gabriel [0000-0002-0606-3622]; Emilie Mathian; Lise Mangiante [0000-0001-8309-0950]; Catherine Voegele; Vincent Cahais [0000-0001-5530-4368]; Akram Ghantous [0000-0002-2582-6402]; James D McKay [0000-0002-1787-3874]; Nicolas Alcala [0000-0002-5961-5064]; Lynnette FernandezCuesta [0000-0002-0724-6703]; Matthieu Foll [0000-0001-9006-8436]
[†]Contributed equally.
[‡]Jointly supervised.

Abstract

Background Lung neuroendocrine neoplasms (NENs) are rare solid cancers, with most genomic studies including a limited number of samples. Recently, generating the first multi-omic dataset for atypical pulmonary carcinoids and the first methylation dataset for large-cell neuroendocrine carcinomas (LCNEC) led us to the discovery of clinically relevant molecular groups as well as a new entity of pulmonary carcinoids (supra-carcinoids). **Results** In order to promote the integration of lung NENs molecular data, we provide here detailed information on data generation and quality control for whole genome/exome sequencing, RNA sequencing, and EPIC 850k methylation arrays for a total of 84 lung NENs patients. We integrate the transcriptomic data with other previously published data and generate the first comprehensive molecular map of lung NENs using the Uniform Manifold Approximation and Projection (UMAP) dimension reduction technique. We show that this map captures the main biological findings of previous studies and can be used as reference to integrate datasets for which RNA sequencing is available. The generated map can be interactively explored and interrogated on the UCSC TumorMap portal (https://tumormap.ucsc.edu/?p=RCG_lungNENomics/LNEN). The data, source code, and compute environments used to generate and evaluate the map as well as the raw data are available respectively in a Nextjournal interactive notebook (https://nextjournal.com/rarecancersgenomics/a-molecular-map-of-lung-neuroendocrine-neoplasms/), and at the EMBL-EBI European Genome-phenome Archive and Gene Expression Omnibus data repositories. **Conclusions** We provide data and all resources needed to integrate it with future lung NENs transcriptomic studies, allowing to draw meaningful conclusions that will eventually lead to a better understanding of this rare understudied disease.

Key words: Carcinoids, lung cancer, neuroendocrine neoplasms, rare cancers, genomics, Tumormap, lungNENomics project

Background

Lung neuroendocrine neoplasms (lung NENs or LNENs) are rare understudied diseases with limited therapeutic opportunities. Lung NENs include poorly differentiated and highly aggressive lung neuroendocrine carcinomas (NECs)–i.e., small-cell lung cancer (SCLC) and large-cell neuroendocrine carcinoma (LCNEC)–as well as well-differentiated and less aggressive lung neuroendocrine tumors (NETs)–i.e., typical and atyp-

Compiled on: September 23, 2020.
Draft manuscript prepared by the author.

ical carcinoids (WHO classification 2015 [1]). Over the past years several genomic studies have investigated the molecular characteristics of these diseases in order to provide some evidence for a more personalized clinical management [2, 3, 4, 5, 6, 7, 8]. Although lung NECs and NETs are broadly considered as different diseases, several recent studies have suggested that they may share some molecular characteristics [9, 10, 7, 11, 12]. However, due to the rarity of these diseases, the sample sizes of these studies individually are limited, and the integration of independent datasets is not an easy task.

Providing a way to interactively visualize and analyze these pan-LNEN data would be of great interest for the scientific community, not only to further explore the proposed molecular link between lung NECs and NETs, but also to integrate data from studies including fewer samples to reach the statistical power needed to draw meaningful conclusions.

Data Description

Recently [7], we performed the first integrative and comparative genomic analysis of lung NEN samples from all histological types, based on newly sequenced data: whole-exome data (WES, 16 samples), whole-genome data (WGS, 3 samples), RNA-Seq data (20 samples), and EPIC 850K methylation data (76 samples), as well as publicly available data. These data correspond to the most extensive multi-omic dataset of lung NENs, including the first methylation data for LCNEC and the first molecular characterization of the rarest lung NEN subtype (atypical carcinoids) [7]. This dataset, which provides the missing pieces for a complete molecular characterization of lung NENs, have been deposited at the EMBL-EBI European Genome-phenome Archive (EGA accession number EGAS00001003699). In order to facilitate the reuse of the data generated in the previous manuscript [7], we provide here a complementary data descriptor by outlining the preprocessing and the quality control (QC) steps performed on each omic dataset available on EGA.

Also, other studies have generated sequencing data and performed a molecular characterization of lung NEN samples: pulmonary carcinoids (mostly typical carinoids) have been characterized by Fernandez-Cuesta *et al.* and Laddha *et al.* [4, 8], LCNEC by George *et al.* [6] and SCLC by George *et al.* [5] and Peifer *et al.* [2]. We therefore generate the first pan-LNEN molecular tumor map by integrating the transcriptomic data from Alcala *et al.* [7] and the other published lung NEN transcriptomic data [2, 4, 5, 6, 8]. This map provides an interactive way to explore the molecular data and allows statistical interrogation, based on the UCSC TumorMap portal [13]. The integrated transcriptomic dataset resulting from these studies is available on GitHub [14].

Data quality controls

Figure 1 provides a schematic view of the preprocessing steps and the associated quality controls performed for each omic dataset generated by Alcala and colleagues [7]. An overview of the available omics and clinical data for each sample is provided in Supplementary Table 1.

WES and WGS data

WES and WGS were performed respectively on 16 and 3 fresh frozen atypical carcinoids in the Cologne Centre for Genomics and the Centre National de Recherche en Génomique Humaine (CNRGH). For WES, the SeqCap EZ v2 Library capture kit from NimbleGen (44Mb) and the Illumina HiSeq 2000 machine (Illumina Inc., CA, USA) were used for the sequencing. For WGS, the Illumina TruSeq DNA PCR-Free Library Preparation Kit was used for library preparation and the HiSeqX5 platform from Illumina for the sequencing as described in [7]. The sequencing reads from the 16 atypical carcinoids whole-exomes and the 3 carcinoids whole-genomes were processed using the in-house Nextflow [15] workflow available at IARCbioinfo/alignment-nf [16] GitHub repository, revision number 9092214665. The pipeline consists in three steps: mapping reads to the reference genome (GRCh37), marking duplicates and sorting reads using bwa v0.7.12-r1044 (RRID:SCR_010910) [17], samblaster v0.1.22 (RRID:SCR_000468) [18], and sambamba v0.5.9 [19] respectively. For WES samples, local realignment using ABRA v0.97b (RRID:SCR_003277) [20] was then run.

The quality controls of the WES and WGS data were performed using FastQC v0.11.8 (RRID:SCR_014583) [21] and QualiMap v2.2.1 (RRID:SCR_001209) [22] using the in-house Nextflow [15] workflows available at IARCbioinfo/fastqc-nf [23] and IARCbioinfo/qualimap-nf [24] repositories respectively, and the results aggregated using MultiQC v1.7 (RRID:SCR_014982) [25] (Figure 1, left panel).

Figure 2A-B, show the per base sequence quality scores (left panels) and the per sequence mean quality scores (right panels). Regarding the per base sequence quality scores, the majority of the base calls were of very good quality (>28, green area, Figure 2A left panel) and of reasonable quality (>20, orange area, Figure 2B left panel) for WES and WGS data respectively. The most frequently observed sequence mean quality score was around 30 for both techniques, which is equivalent to an error probability of 0.1%. Table 1 provides the general statistics associated to the WES and WGS quality controls. The observed median coverage for each sample was above the expected coverage (30X for the WGS samples and 120X for the WES samples). Concerning the alignment quality, all WES samples had more than 99% of the reads aligned and all WGS samples had more than 98% of the reads aligned.

RNA-Seq data

RNA-Sequencing was performed on 20 fresh frozen atypical samples. The Illumina TruSeq RNA sample preparation Kit was used for library preparation and the Illumina TruSeq PE Cluster Kit v3 and the Illumina TruSeq SBS Kit v3-HS kits were used on an Illumina HiSeq 2000 sequencer. The data generated were processed in five steps (Figure 1, middle panel): i) reads trimming using Trim Galore v0.6.5 (RRID:SCR_011847) [26], ii) reads mapping to the reference genome (GRCh38, gencode version 33 from bundle CTAT from 6th April 2020 [27]) using STAR v.2.7.3a (RRID:SCR_015899) [28], iii) realignment of the reads using ABRA2 v2.22 (RRID:SCR_003277) [29], iv) base quality score recalibration using GATK4 v4.0.5.1 (RRID:SCR_001876) [30, 31] and v) gene expression quantification using StringTie v2.1.1 (RRID:SCR_016323) [32]. FastQC v.0.11.9 (RRID:SCR_014583) [21], RSeQC v3.0.1 (RRID:SCR_005275) [33] and HTSeq v0.12.4 (RRID:SCR_005514) [34] were used to control the raw reads quality and assignments, and the results aggregated using MultiQC v1.7 (RRID:SCR_014982) [25]. These steps were performed using our in-house Nextflow [15] pipelines available at the following GitHub repositories: IARCbioinfo/RNAseq-nf [35] release v2.3, IARCbioinfo/abra-nf [36] release v3.0, IARCbioinfo/BQSR-nf [37] release v1.1 and IARCbioinfo/RNAseq-transcript-nf [38] release v2.1.

Figure 2C shows that the base calls, before trimming, are of good quality since all samples have a mean per base sequence quality score higher than 28 (left panel) and for all samples the most frequently observed per sequence mean qual-

Table 1. General statistics associated to the quality controls of the WES and WGS data

Sample	Sequencing	Median coverage	Total nb reads (M)	>30x (%)	Aligned (%)	GC (%)	Median insert size	Duplicates (%)
LNEN002	WES	148	113.3	95.5	99.7	53.7	194	13.9
LNEN003	WES	146	110.3	95.8	99.7	53.7	194	13.4
LNEN004	WES	150	115.3	95.4	99.8	54.3	193	13.1
LNEN005	WES	135	103.4	94.7	99.8	54	195	12.1
LNEN006	WES	126	93.6	94.6	99.8	53.5	197	12.5
LNEN007	WES	145	116.3	94.4	99.8	54.5	195	14.8
LNEN009	WES	123	98.4	92.9	99.7	54.1	195	12.4
LNEN010	WES	138	104.1	95	99.7	53.3	196	13.4
LNEN011	WES	161	125.8	95.8	99.8	54.3	196	14.8
LNEN013	WES	131	99.2	94.3	99.8	53.5	193	13
LNEN014	WES	132	102.6	94	99.8	54.1	195	13.3
LNEN015	WES	148	111.3	95.7	99.6	54.1	197	10.1
LNEN016	WES	133	98	94.3	99.6	54.3	194	9
LNEN017	WES	158	116.4	95.9	99.6	54.1	192	8.9
LNEN020	WES	187	144.7	96.6	99.7	53.6	192	14.5
S00716_B	WES	133	99.8	95.4	99.7	52.8	194	14.3
LNEN041	WGS	36	923.5	77.5	98.9	41	366	13.3
LNEN042	WGS	41	993.7	88.1	98.8	41.5	388	9.4
LNEN043	WGS	43	1033.1	89.7	99.3	41.6	392	8.8

ity is above 35, corresponding to an error probability of 0.03%, (right panel). None of the samples presented more than 1% of over-represented sequences, which assures a proper library diversity. RSeQC was used to control the alignment quality and to assign mapped reads to different genomic features (coding regions, introns, intergenic regions, TSS, TES). Figure 2D (left panel) shows that every sample had more than 70% of reads uniquely mapped and the reads distribution for each sample is represented on Figure 2D (middle panel). All samples had more than 75% reads mapped in coding regions (CDS-exons, 5' and 3' UTR exons). The reads counting was performed at the gene level for 59,607 genes (genecode annotation, release 33) using HTSeq [34]. Figure 2D (right panel) shows the reads assignments, the percentage of assigned reads ranges from 71.3 to 87.3%. STAR, RSeQC and HTSeq metrics for each sample are provided in Supplementary Tables 2-4. Note that three samples, LNEN008, LNEN014 and LNEN017, have a higher proportion of reads classified as "Unmapped too short" and "Mapped to multiple loci" (Figure 2D, left panel), reads mapped in intronic regions (Figure 2D, middle panel) and a lower proportion of reads assigned by HTSeq (Figure 2D, right panel) in comparison to the other samples. Unexpected results concerning those samples should be thus considered with caution.

Finally, in order to apply dimensionality reduction methods to the RNA-Seq data (see below), the DESeq2 package v1.26.0 (RRID:SCR_015687) [39] was used to transform the read counts obtained using StringTie to variance stabilized read counts (vst), enabling the comparison of samples with different library sizes. To reduce sex influence on expression profiles, the genes located on sex chromosomes were not considered for subsequent analyses. Genes located on mitochondria chromosomes were as well not considered.

Methylation data

The methylation analyses were performed based on the EPIC 850K methylation arrays and the Infinium EPIC DNA methylation beadchip platform (Illumina) for 33 typical carcinoids, 23 atypical carcinoids, 20 LCNEC and 19 technical replicates in total. These arrays interrogate more than 850,000 CpGs and contain internal control probes that can be used to assess the overall efficiency of the sample preparation steps. The raw intensity data (IDAT files) were processed using the R package *minfi* v.1.24.0 (RRID:SCR_012830) [40]. Figure 1 (right panel) provides the packages, functions and publication used for the data processing, quality control and filtering steps as implemented in the IARCbioinfo/Methylation_analysis_scripts [41] GitHub repository.

Figure 2E shows that no outliers were detected: i) the left panel, representing the median log2 of the methylated and un-

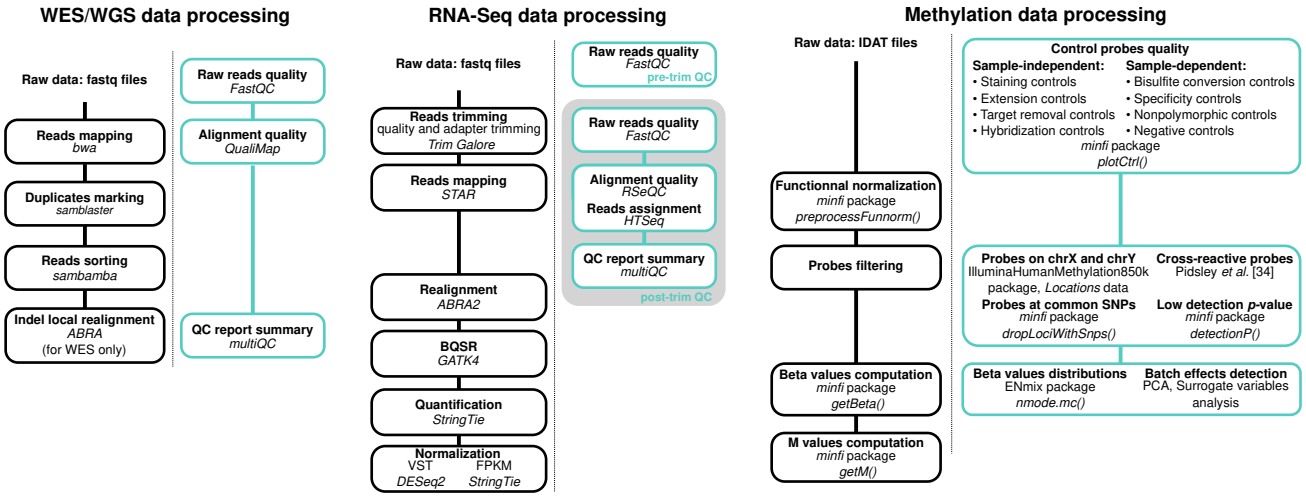


Figure 1. Bioinformatics workflows for data processing and associated quality controls. Bioinformatics tools used for the processing of the WES/WGS data, RNA-Seq and methylation data are represented in the left, middle and right panels respectively. Green boxes correspond to quality controls (QC) steps.

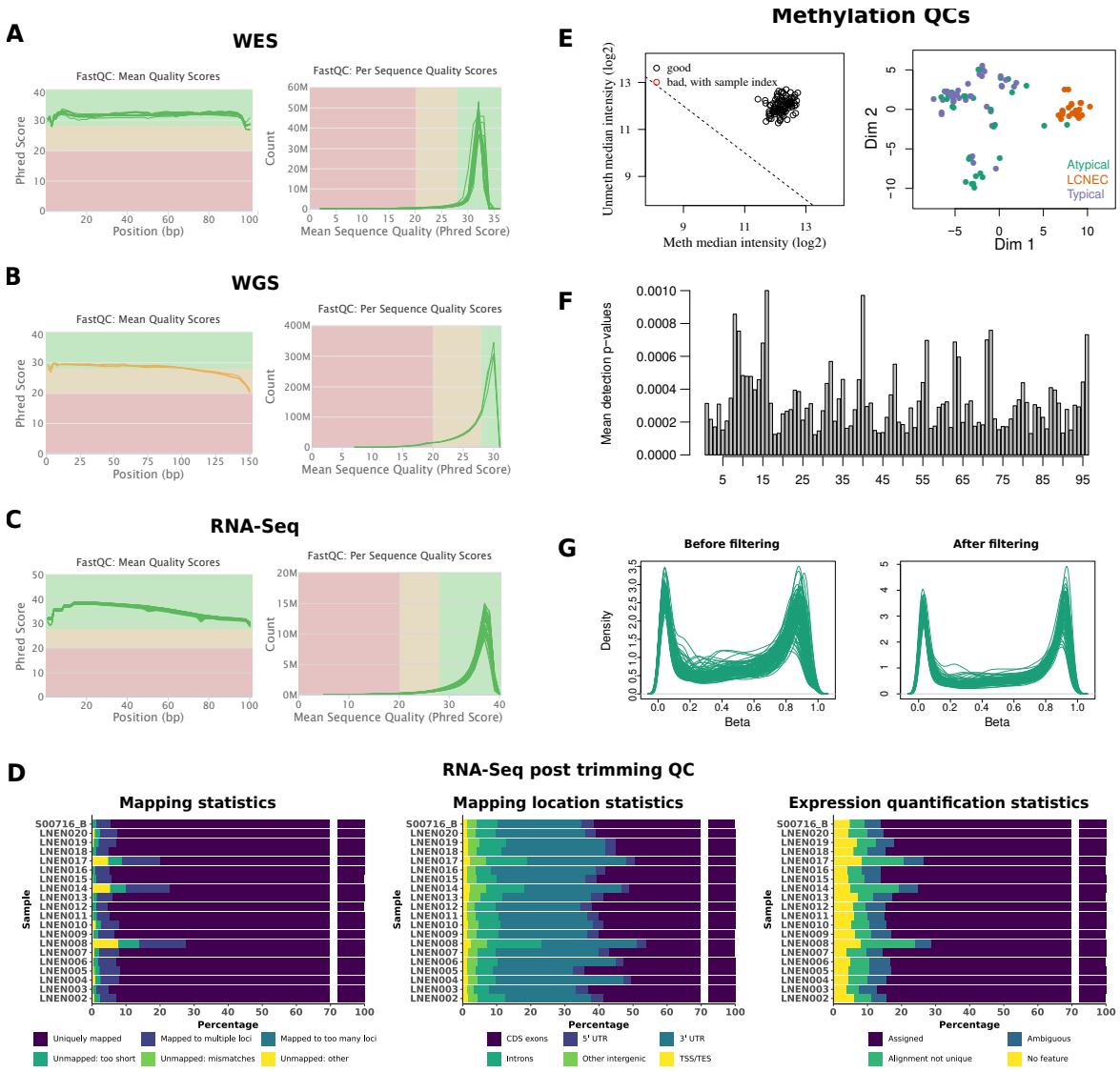


Figure 2. Quality controls performed on each omic dataset. A) Reads quality control using FastQC for WES data. **B)** Reads quality control using FastQC for WGS data. **C)** Reads quality control using FastQC for RNA-Seq data. For A, B, and C, the left panels correspond to the sequence quality plots, the x-axis representing the base position in the read and the y-axis the mean quality value; the right panels correspond to the per sequence quality scores plots, the x-axis representing the mean quality score and the y-axis the number of reads. **D)** Quality control of the RNA-Seq data after trimming. Left panel: barplot representing the percentages of reads uniquely mapped ("Uniquely mapped"), mapped to multiple loci ("Mapped to multiple loci" or "Mapped to too many loci" if the number of loci is higher than 10), unmapped because the mapped reads' proportion was too small ("Unmapped: too short"), unmapped because of too many mismatches ("Unmapped: mismatches"), or unmapped for other reasons ("Unmapped: other"). Middle panel: cumulative barplot representing the percentages of reads mapped, using RSeQC, at different locations in the genome (exons, introns, 5' and 3' UTR, intergenic regions, TSS, and TES). Right panel: cumulative barplot representing the cumulative percentages associated to the different reads assignments using HTSeq ("Assigned": reads assigned to one gene, "Ambiguous": reads assigned to multiple overlapping genes, "Aligned not unique": reads assigned to multiple non-overlapping genes, "No Feature": reads assigned to none of the features). **E)** Left panel: samples' quality based on log median intensities. The x-axis and y-axis correspond to the median of log2 methylated and unmethylated intensities, respectively. Right panel: representation of the between-sample similarities based on the two first MDS dimensions. **F)** Histogram of the median detection p-value for each sample. **G)** Distribution of the beta values for each sample before and after the filtering step (left and right panel respectively).

methylated intensities, indicates that all samples cluster together with a log median intensity above 11 for both channels, which supports the absence of failed samples, ii) on the right panel, the multidimensional scaling (MDS) plot shows that the samples cluster together by histological groups. We used the *depectionP* function (*minfi* package), which compares the DNA signal to the background signal based on the negative control probes to provide a detection *p-value* per probe, lower *p-value* indicating reliable CpGs. Figure 2F represents the mean detection *p-values* per sample and shows that all samples mean detection *p-values* were lower than 0.01. To correct for the variability identified in the control probes, a normalization step was applied to the raw intensities using the *preprocessFunnorm* function from *minfi*.

After between-array normalization, different sets of probes that could generate artefacts were removed successively from the methylation dataset: i) 19634 probes on the sex chromosomes, in order to identify differences related to tumors but unrelated to sex chromosomes, ii) 41818 cross-reactive probes which are probes co-hybridizing with multiple CpGs on the genome and not only to the one it has been designed for [42], iii) 10588 probes associated with common SNPs (present in dbSNP build 137), iv) 24363 probes with multi-modal beta-value distribution, and v) 9697 probes having a detection *p*-value higher than 0.01 in at least one sample. Supplementary Table 5 lists the sets of filtered probes. To assess the experimental quality of the assay, the distributions of the beta values were analyzed. As described previously, probes with multi-modal

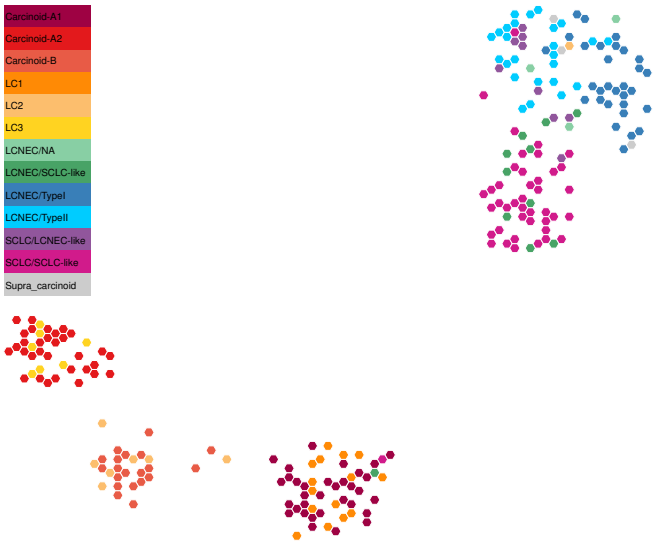


Figure 3. Two dimensional projection of lung NENs transcriptome data using UMAP. The representation was obtained from the TumorMap portal, using the hexagonal grid view, each hexagonal point representing a lung NEN sample. Point colors correspond to the molecular clusters defined in the previous manuscripts.

distributions were removed at the filtering step and overall distributions of beta values for each sample before and after filtering were plotted (Figure 2G). As expected, after filtering all samples showed a bimodal profile, indicative of the good quality of the experiment. No experimental batch effects were identified after functional normalization (see Supplementary Fig. 33 from [7]). Based on all the quality controls performed, none of the samples analyzed were identified as outlier. However, one sample available on EGA (201414140007_R06C01), was removed from the analyses because it came from a metastatic tumor rather than the primary tumor. Samples metadata are provided in Supplementary Table 6.

Generation of an integrative molecular map

Here we have generated a pan-LNEN molecular map with the whole-transcriptomic (RNA-Seq) data available from individual studies of each lung NEN tumor type [2, 4, 5, 6, 7, 8]. This dataset includes the RNA-Seq data for a total of 51 SCLC, 69 LCNEC, 118 carcinoids including 40 atypical and 75 typical carcinoids. The different data underwent the same processing steps described above since the generation of the molecular map requires a homogenized dataset.

Dimensionality reduction using UMAP

UMAP method

The pan-LNEN map was obtained using the Uniform Manifold Approximation and Projection (UMAP) method [43] on the genes with the most variable expression (genes explaining 50% of the total variance). UMAP is a dimensionality reduction method based on manifold learning techniques, which are adapted to non-linear data in contrast with the commonly used PCA method. Firstly, it builds a topological representation of the high-dimensional data, and secondly it finds the best low-dimensional representation of this topological structure [43]. UMAP representations were generated using the umap function from the R package umap (v. 0.2.5.0) [44]. All the parameters were set to their default values except the *n_neighbors* parameter. This parameter defines the number of neighbors considered to learn the structure of the topological space. Varying this parameter from small to large values enables the user to find a trade-off between local and global preservation of the space, respectively. In order to preserve the global structure of the data (see "quality control of the UMAP projection" section below), we built the pan-LNEN map setting the *n_neighbors* parameter to 238, which corresponds to the total number of samples.

Biological interpretation of the pan-LNEN TumorMap

Figure 3 shows the pan-LNEN map available on TumorMap [45] (see "Re-use potential" section below), with colors representing the main molecular subtypes. To evaluate the accuracy of the generated pan-LNEN map we firstly verified whether it was consistent with the main biological findings from the original studies, in particular whether it represented the molecular subtypes of lung NENs previously identified, and their relationship with histological types. We specifically tested whether groups of samples previously described as having discordant molecular and histopathological features were identified in our map. To do so, given a focal molecular subtype and two reference histopathological types, we assessed whether samples from the focal molecular subtype were closer to one of the two references using a one-sided Wilcoxon test between the euclidean distances of samples to the centroid of each reference type.

First, the SCLC/LCNEC-like samples [6], which are histological SCLCs presenting the molecular profile of LCNEC, tend to cluster with the LCNECs rather than with the SCLCs (Wilcoxon $p\text{-}value = 6.2 \times 10^{-4}$). Similarly, the LCNEC/SCLC-like samples [6], which are histological LCNECs having the molecular profile of SCLC, tend to cluster with the SCLCs rather than with the LCNECs (Wilcoxon $p\text{-}value = 3.3 \times 10^{-3}$). In 2018, George *et al.* showed also that LCNEC samples can be subdivided into the type-I and type-II molecular groups [6]. We observed that the type-I and type-II LCNECs were closer to each other than to the SCLC/SCLC-like (Wilcoxon $p\text{-}value = 9.9 \times 10^{-14}$) and that SCLC/LCNEC-like samples were closer to type-II than type-I LCNECs [6] (Wilcoxon $p\text{-}value = 3.9 \times 10^{-3}$). Like the LCNECs, pulmonary carcinoids have been subdivided in molecular groups. Alcala *et al.* [7] identified three clinically relevant molecular clusters, using a multi-omics factor analysis (MOFA): Carcinoid A1, Carcinoid A2, and Carcinoid B [7]. In the pan-LNEN map generated using UMAP, those three clusters are clearly visible (Figure 3) and respectively correspond to the three clusters identified in [8] named LC1, LC3 and LC2. Also, in the study from Alcala and colleagues [7], two carcinoids that

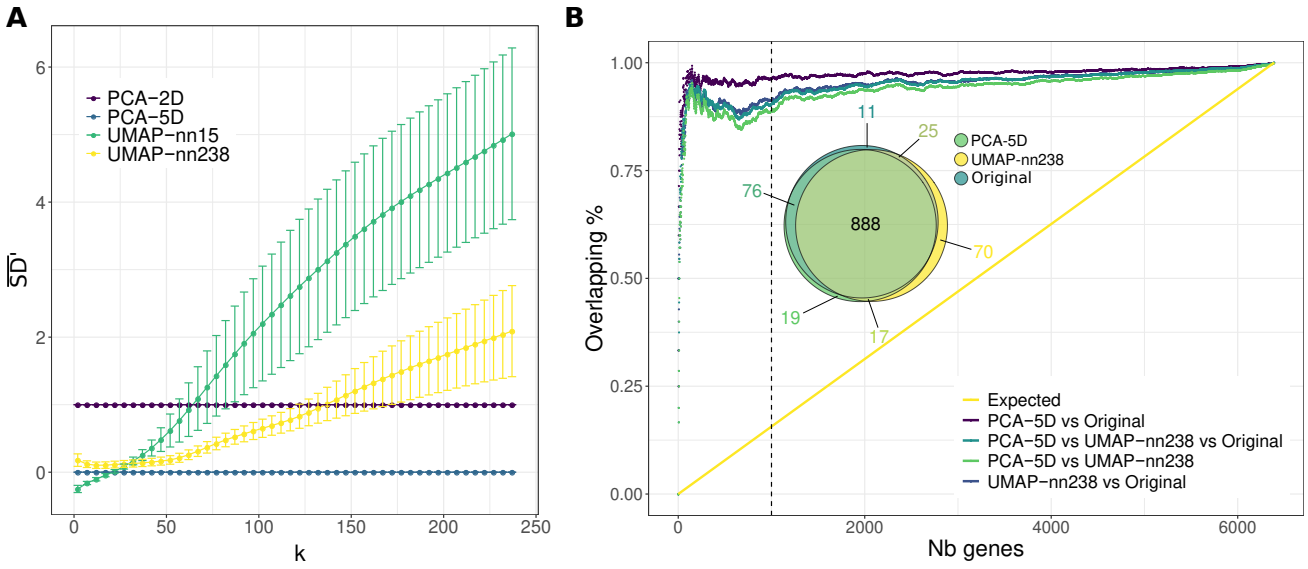


Figure 4. Quality controls performed on the UMAP projection. A) Comparison of the samples' neighborhood preservation for UMAP, PCA-2D, and PCA-5D dimensionality reductions. $\overline{SD}'_k$ values are represented as a function of the number k of nearest neighbors considered, for different dimensionality reduction methods: PCA-2D in purple, PCA-5D in blue, UMAP with *n_neighbors* = 238 (UMAP-nn-238) in yellow and UMAP with the default value *n_neighbors* = 15 (UMAP-nn-15) in green. Error bars correspond to the means more or less the standard deviations computed across 1000 replicate simulations. **B)** Concordance between gene expressions' spatial auto-correlations in the original space, UMAP-nn-238, and PCA-5D dimensionality reductions. For each space, the genes were ranked based on the spatial auto-correlations of their expression (mean MI values). The concordance is measured as the proportion of overlap between the top N genes in the different spaces (colored lines). The yellow line corresponds to the proportion of overlap expected under the null hypothesis (based on the expected mean of the hypergeometric law). The Euler diagram represents the overlaps between the top 1000 features (N = 1000, dashed line) resulting from the three spaces.

clustered with the carcinoids B (S00118 and S00089) were borderline and located between cluster A1 and B. Similarly, a LCNEC sample and a SCLC sample clustered with the carcinoids A1 [7]. These observations are also visible on the TumorMap representation. Finally, in the same study, a novel entity of carcinoids, named the supra-carcinoids was unveiled. These samples were characterized by a morphology similar to that of pulmonary carcinoids but the molecular features of LCNEC samples. In the pan-LNEN TumorMap, the supra-carcinoids also clustered with the LCNEC samples and were molecularly closer to LCNECs than to SCLCs (Wilcoxon $p\text{-}value = 5 \times 10^{-2}$). We also note that one sample from Laddha *et al.* [8] LC2 cluster (SRR7646258) clusters with LCNEC.

Quality control of the UMAP projection

In any dimensional reduction technique, there is a trade-off between preserving the global structure of the data and the fine scale details, and UMAP has been designed to reach a better balance compared to previous methods.

Based on the previously published analyses of lung NEN data [2, 4, 5, 6, 7, 8], we expect the global structure of the data to be composed of six molecular groups (SCLCs, type I and type II LCNECs, Carcinoid A1, A2 and B). For this reason, an ideal projection able to capture this large scale variation should contain five dimensions. To assess the quality of the 2-dimensional representation generated by UMAP, we propose a comparative analysis between UMAP and the traditional principal component analysis (PCA) based on the five first principal components of PCA (PCA-5D) as implemented in the *dudi.pca* function from the ade4 R package (v1.7-15) [46]. Because UMAP is aiming at preserving the global structure in only two dimensions, we also compared it to the traditional PCA based only on the two first principal components (PCA-2D). We evaluated the performance of the methods based on the preservation of: (i) the samples' neighborhood and (ii) the spatial auto-correlations.

Preservation of the samples' neighborhood

We used the sequence difference view (SD) metric (eq. 3 from [47]) to evaluate the preservation of the samples' neighborhood. This dissimilarity metric compares, for a given sample, its neighborhood in the low-dimensional space with that in the original space, taking into account that preserving the rank of a close neighbor is more important than for a distant neighbor (see [47] for details). SD values are positive ($SD \in [0\,;+\infty)$), with small values indicating a good preservation of the samples neighborhood. We denote by $\overline{SD}_k$ the value of SD averaged across samples for a fixed number of neighbors k; $\overline{SD}_k$ gives a sense of the overall preservation of the neighborhood at different scales: local for low k values and global for large k values. We calculated $\overline{SD}_k$ for PCA-5D, PCA-2D, UMAP with *n_neighbors* = 238 and UMAP with the default value *n_neighbors* = 15. Because we are interested in the relative values of $\overline{SD}_k$ for the different dimensionality reduction methods, and because we use PCA as a reference, for each dimensionality reduction method X we scaled the values of $\overline{SD}_k$ using that of PCA-5D and PCA-2D:

$$\overline{SD}'_{k,X} = \frac{\overline{SD}_{k,X} - \overline{SD}_{k,PCA-5D}}{\overline{SD}_{k,PCA-2D} - \overline{SD}_{k,PCA-5D}}. \quad (1)$$

By definition, $\overline{SD}'_{k,PCA-5D} = 0$ and $\overline{SD}'_{k,PCA-2D} = 1$. Thus values of $\overline{SD}'_{k,X}$ close to 0 indicate that X preserves k neighborhoods as well as PCA-5D, whereas values close to 1 indicate that X preserves k neighborhoods worse than PCA-5D but as well as PCA-2D, and values greater than 1 indicate that X preserves k neighborhoods worse than PCA-2D and PCA-5D. Note that $\overline{SD}'_{k,X}$ can be negative if X preserves k neighborhoods better than $\overline{SD}_{k,PCA-5D}$. For the UMAP projection, we iterated the computation of $\overline{SD}'_k$ 1000 times, because the algorithm uses a stochastic optimization step to define the projection.

As expected, increasing the *n_neighbors* UMAP parameter from 15 to 238 leads to a better preservation of the global struc-

ture, clearly visible for $k > 30$ (Figure 4A; mean $\overline{SD}'_{k>30}$ equals to 2.855 and 1.029 respectively), while only marginally reducing the preservation of the local structure for $k < 30$ (mean $\overline{SD}'_{k<30}$ equals to -0.076 and 0.124 respectively), which is approximately the size of the smallest cluster. Globally, the $\overline{SD}'_k$ values over all k levels are lower for a *n_neighbors* value of 238 than 15 (paired t-test *p-value* = 6.09×10^{-8}). With *n_neighbors* = 238, the UMAP projection provides a clear improvement over PCA-2D for k around 135 (mean $\overline{SD}'_k < 1$), offering a good trade-off for visualisation in only two dimensions while being able to maintain the global structure of the data, in particular the six molecular groups previously identified. This observation highlights the importance of varying the *n_neighbors* parameter according to the purpose of the projection. Some analyses would require to maintain the local structure of the samples neighborhood while others the global structure.

Preservation of spatial auto-correlations

Under the hypothesis that close points on projections share a similar molecular profile, spatial auto-correlations were measured according to the Moran Index (MI) metric [48]. MI values range from -1 to 1, the extreme values indicating negative (nearby locations have dissimilar gene expression) or positive (nearby locations have similar gene expression) spatial auto-correlation, respectively. The spatial auto-correlation of the expression of each gene helps to identify the genes contributing to the structure of the molecular map (MI$\simeq$ 1), and conversely, the genes that are randomly distributed spatially (MI$\simeq$ 0). The computation of MI requires a weight matrix that determines the spatial scale at which auto-correlation is assessed; we gave a weight of 1 to the k nearest neighbors based on Euclidean distance, and 0 otherwise, so that we can control the scale at which MI is computed with parameter k. The mean MI across k values was computed for all gene expression features for: (i) the original space, (ii) the PCA-5D projection, and (iii) the UMAP projection (with *n_neighbors* = 238). We used the implementation of MI from the Moran.I function of R package ape (v. 5.3) [49].

To evaluate the preservation of the spatial auto-correlations, we ranked the top N genes based on the mean MI values for these three cases and calculated the overlap between the lists (Figure 4B). We found that the PCA-5D is only slightly more conservative of the spatial auto-correlations found in the original space than UMAP (unilateral paired t-test *p.value* = 2.2×10^{-16}). For example, for N = 1000 (see Euler diagram inserted in Figure 4B), 88.8% of the genes with the highest MI overlap between the PCA-5D, UMAP and the original space.

Re-use potential

An interactive TumorMap

Newton and colleagues have recently developed a portal called TumorMap [13, 50], an online tool dedicated to omics data visualization. This new type of integrated genomics portal uses the Google Maps technology designed to facilitate visualization, exploration, and basic statistical interrogation of high dimensional and complex datasets. The pan-LNEN molecular map that we generated in this work (Figure 3) has been shared on the TumorMap platform. Along with the molecular map, the main clinical, histopathological and molecular features highlighted in the previous studies were uploaded as attributes. The interface enables users to explore and navigate through the map: zooming in and out, coloring and filtering samples based on attributes. The users can also create their own attributes based on pre-existing ones by using operators such as union or intersection. In addition, multiple statistical tests are pre-implemented and available, for example: comparison of attributes without considering the samples positions on the map, comparison of attributes considering samples positions on the map, and ordering attributes based on their potential to differentiate two groups of samples. The interactive nature of the map and the fact that its manipulation does not require computational expertise, could enable the generation of new hypotheses and expand the reuse potential of the dataset.

An interactive computational notebook

In the first part of the paper, we described the pre-processing and quality control steps applied on the recently published lung NEN multi-omics dataset [7] in order to facilitate its reuse. To generate the pan-LNEN molecular map, the same pre-processing steps were followed to homogenize independently published transcriptomic data [2, 4, 5, 6, 7, 8]. For that purpose, reproducible pipelines, developed in house, were used and are available for reuse to the scientific community on GitHub [51] (see the "availability of source code" section). In addition, the code used to generate the molecular map and to evaluate the quality of the dimensionality reduction is provided as a notebook published on Nextjournal [52]. Along with the code, the notebook provides the data and the dependencies required to run the analyses performed in this paper. Interested researchers can thus make a copy of this publicly available notebook (called "Remix") to reproduce our results but also interactively modify the code and explore the influence of different parameters.

Integration of new samples

The homogenized read counts of the pan-LNEN data are available on GitHub [14]. Along with the available code, these data could be used to integrate new samples for which RNA-Seq data are available. The raw read counts of the new samples should firstly be generated following the same processing steps described in the section "Data quality controls" (Figure 1, middle panel) and integrated to the pan-LNEN read counts. We also provide in the Nextjournal notebook, the Nextflow command lines allowing to obtain the read counts. The variance stabilized transformation (DESeq2 [39]) should then be applied on the combined data set and UMAP should finally be rerun to project all samples together in a two dimensional space. All together, we provide the resources to integrate additional samples into our molecular map, starting from raw sequencing read counts.

Discussion

Genomic projects focused on rare cancers encounter the limitation of availability of good quality biological material suitable for such studies. This translates in small series of samples usually underpowered to draw meaningful conclusions. Thus, tools facilitating the integration of independent datasets into larger sample series will lead to more informative studies. Recently, the first multi-omic dataset for the understudied atypical pulmonary carcinoids and the first methylation dataset for LCNECs was published [7]. Here we provide a parallel description of the pre-processing of these molecular data and provide evidence of the good quality of the different 'omics data generated. This data collection associated with previous datasets [2, 4, 5, 6, 8] completes the lung NENs molecular landscape and provides thus a valuable resource to improve the molecular characterization of lung NEN tumors. Notably, we show

here the perfect concordance of the three molecular clusters of pulmonary carcinoids independently identified in [7] and [8], validating the discoveries made by these two studies and proving the usefulness of this integrative approach.

However, even when primary genomic data is available, barriers to accessing the data still exist, often limiting its reuse by the community [53]. In particular, downloading and reprocessing large raw sequencing data requires dedicated infrastructure and bioinformatics skills. Indeed, in order to minimize batch effects when integrating data from different studies, one need to process it exactly in the same way (with the same software and the same versions, the same reference genome, the same annotation databases *etc.*). As more and more data are generated, the previously mentioned reprocessing will become untenable and replicating these efforts for each new study in each research group represents a waste of resources. Standardization of laboratory and computational protocols might become a reality when large national medical genomics initiatives will be fully operational [54]. In the meantime there is a need for better data sharing strategies than the traditional "supplementary spreadsheet / raw data" combination that can accelerate the translational impact of molecular findings.

One step in this direction is the generation of so called "tumor maps", which provide an interactive way to explore the molecular data and allow easy statistical interrogation, including generating new hypotheses, but also projecting data from future studies including fewer samples [13]. This integration method has some limitations though. A fixed reference map could be of interest for easier biological interpretations, but the overall sample size of the datasets used to build the pan-LNEN map remains relatively small. Thus, the map does probably not capture the complete molecular diversity of the lung NENs, and integrating new samples will influence the map and potentially change the clusters obtained after dimensionality reduction. Also, if the harmonization of the new dataset to integrate is not enough to correct for strong batch effects, the interpretation of the projections would be erroneous. Another approach would be to project the new samples into a fixed reference map. However, the stochastic nature of UMAP embedding and its sensibility to parameter tuning can lead to unstable projection results, thus this task is for now not straightforward and requires further development [55]. In the meantime, favoring the integration of datasets will, over the years, yield to the constitution of molecular maps that will probably be more and more accurate and more adapted to the projection of new samples.

Conclusion

Here we provide a molecular map based on homogenized transcriptomic data available for the four types of lung NENs from six different studies. We show that this map represents well both the local and global structure of the data, and captures the main biological features previously reported. We provide a full spectrum of data and tools to maximize its re-use potential for a wide range of users: raw sequencing reads, gene expression matrix, bioinformatics pipelines, interactive computational notebooks and an interactive TumorMap. In particular, we indicate how one can update the molecular map by integrating new samples starting from raw sequencing reads. Considering the small sample sizes of molecular studies on rare lung NENs, promoting data integration will empower more reliable statistical testing, and this map will therefore serve as a reference in future studies.

Availability of supporting data and materials

R codes used for this article are available in the GigaDB data repository [56]. The data used in this manuscript are available on the European Genome-phenome Archive (EGA) which is hosted at the EBI and the CRG, under the accession numbers EGAS00001003699, EGAS00001000650, EGAS00001000925, EGAS00001000708, as well as on Gene expression Omnibus (GEO) under GEO SuperSeries GSE118131.

Declarations

Ethical Approval

These data belong to the lungNENomics project, which has been approved by the IARC Ethical Committee.

Consent for publication

Not applicable.

Competing Interests

The authors declare no conflict of interest. Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Funding

This work has been funded by the US National Institutes of Health (NIH R03CA195253 to L.F.C. and J.D.M.), the French National Cancer Institute (INCa, PRT-K-17-047 to L.F.C. and M.F.), the Ligue Nationale contre le Cancer (LNCC 2016 to L.F.C.), France Genomique (to J.D.M), and the Neuroendocrine Tumor Research Foundation (NETRF, Investigator Award 2019 to L.F.C.). L.M. has a fellowship from the LNCC.

List of abbreviations

Additional files

Supplementary Table 1: Samples overview
Supplementary Table 2: Summary table of STAR metrics
Supplementary Table 3: Summary table of RSeQC metrics
Supplementary Table 4: Summary table of HTSeq metrics
Supplementary Table 5: List of filtered probes
Supplementary Table 6: Samples methylation metadata

Author's Contributions

MF and LFC conceived and designed the study. AAGG, EM, NA, LM and CV performed the analyses. VC and AG gave scientific input for the methylation part. JDM helped with logistics and gave scientific input. AAGG, EM, NA, MF and LFC wrote the manuscript. All the authors read and commented the manuscript.

AC	Atypical carcinoids
ABRA	Assembly-based realigner
BAM	Binary Alignment Map
CDS	Coding Sequence
CGR	Center for Genomic Regulation
CpG	Cytosine–Phosphate–Guanine
CTAT	The Trinity Cancer Transcriptome Analysis Toolkit
dbSNP	The Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
EGA	European Genome-phenome Archive
EMBL-EBI	The European Bioinformatics Institute
GATK	Genome Analysis Toolkit
IDAT	File format of the raw methylation data
LCNEC	Large-cell neuroendocrine carcinoma
LCNEC/SCLC-like	Large-cell neuroendocrine carcinomas with the molecular features of small cell lung cancers
LNEN	Lung neuroendocrine neoplasm
MDS	Multidimensional scaling
MI	Moran's Index
MOFA	Multi-omics factor analysis
NEC	Neuroendocrine carcinomas
NEN	Neuroendocrine neoplasm
NET	Neuroendocrine tumors
PCA	Principal Component Analysis
QC	Quality control
RNA-Seq	Ribonucleic acid sequencing
SCLC	Small-cell lung cancer
SCLC/LCNEC-like	Small cell lung cancers with the molecular features of large-cell neuroendocrine carcinomas
SCLC/SCLC-like	Small cell lung cancers with the molecular features of small cell lung cancers
SD	Sequence Difference view metric
SNP	Single Nucleotide Polymorphism
STAR	Spliced Transcripts Alignment to a Reference
TC	Typical carcinoids
TES	Transcription End Site
TSS	Transcription Start Site
UCSC	University of California Santa Cruz
UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated Transcribed Region
vst	Variance Stabilized Transformation
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization

Acknowledgements

This study is part of the lungNENomics project and the Rare Cancers Genomics initiative (http://rarecancersgenomics.com). We also acknowledge the Cologne Centre for Genomics (Cologne, Germany) and the Centre National de Recherche en Génomique Humaine (Evry, France) for generating good quality sequencing data. We also thank Cyrille Cuenin and Zdenko Herceg from the Epigenetics group at IARC; and Teresa Swatloski and Josh Stuart from UCSC for their assistance in hosting our map on the UCSC tumormap portal.

References

1. Rindi G, Klimstra DS, Abedi-Ardekani B, Asa SL, Bosman FT, Brambilla E, et al. A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. Modern Pathology 2018;31(12):1770–1786.
2. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nature Genetics 2012;44(10):1104–1110.
3. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nature Genetics 2012;44(10):1111–1116.
4. Fernandez-Cuesta L, Peifer M, Lu X, Sun R, Ozretić L, Seidal D, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nature communications 2014;5:3518.
5. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. Nature 2015;524(7563):47–53.
6. George J, Walter V, Peifer M, Alexandrov LB, Seidel D, Leenders F, et al. Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. Nature Communications 2018;9(1):1048.
7. Alcala N, Leblay N, Gabriel AAG, Mangiante L, Hervas D, Giffon T, et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. Nature Communications 2019;10(1):3407.
8. Laddha SV, Da Silva EM, Robzyk K, Untch BR, Ke H, Rekhtman N, et al. Integrative genomic characterization identifies molecular subtypes of lung carcinoids. Cancer Research 2019;79(17):4339–4347.
9. Pelosi G, Bianchi F, Dama E, Simbolo M, Mafficini A, Sonzogni A, et al. Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. Virchows Archiv 2018;472(4):567–577.
10. Rekhtman N, Pietanza MC, Hellmann MD, Naidoo J, Arora A, Won H, et al. Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma–like and Non–Small Cell Carcinoma–like Subsets. Clinical Cancer Research 2016;22(14):3618–3629.
11. Simbolo M, Barbi S, Fassan M, Mafficini A, Ali G, Vicentini C, et al. Gene Expression Profiling of Lung Atypical Carcinoids and Large Cell Neuroendocrine Carcinomas Identifies Three Transcriptomic Subtypes with Specific Genomic Alterations. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2019;14(9):1651–1661.
12. Fernandez-Cuesta L, Foll M. Molecular studies of lung neuroendocrine neoplasms uncover new concepts and entities. Translational Lung Cancer Research 2019;8(S4).
13. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Research 2017;77(21):e111–e114.
14. IARCbioinfo/DRMetrics GitHub repository. https://github.com/IARCbioinfo/DRMetrics, accessed January 2020.
15. Tommaso PD, Floden EW, Magis C, Palumbo E, Notredame C. Nextflow, an efficient tool to improve computation numerical stability in genomic analysis. Biol Aujourdhui 2017;211(3):233–237.
16. IARCbioinfo/alignment-nf GitHub repository. https://github.com/IARCbioinfo/alignment-nf, accessed March 2018.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–1760.
18. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 2014;30(17):2503–5.
19. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31(12):2032–2034.
20. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics 2014;30(19):2813–5.
21. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S, FastQC. Babraham, UK; 2012. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed August 2019.
22. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics

2016;32(2):292–294.
23. IARCbioinfo/fastqc-nf GitHub repository. https://github.com/IARCbioinfo/fastqc-nf, accessed August 2019.
24. IARCbioinfo/qualimap-nf GitHub repository. https://github.com/IARCbioinfo/qualimap-nf, accessed August 2019.
25. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32(19):3047–8.
26. Krueger F, Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries; 2012. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed March 2018.
27. CTAT libraries. https://data.broadinstitute.org/Trinity/CTAT_RESOURCE_LIB/, accessed May 2020.
28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15.
29. Mose LE, Perou CM, Parker JS. Improved indel detection in DNA and RNA via realignment with ABRA2. Bioinformatics 2019 sep;35(17):2966–2973.
30. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 2011 may;43(5):491–501.
31. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics 2013;11(SUPL.43):11.10.1.
32. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 2015;33(3):290–295.
33. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics 2012;28(16):2184–2185.
34. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31(2):166.
35. IARCbioinfo/RNAseq-nf GitHub repository. https://github.com/IARCbioinfo/RNAseq-nf, accessed May 2020.
36. IARCbioinfo/RNAseq-nf GitHub repository. https://github.com/IARCbioinfo/abra-nf, accessed May 2020.
37. IARCbioinfo/RNAseq-nf GitHub repository. https://github.com/IARCbioinfo/BQSR-nf, accessed May 2020.
38. IARCbioinfo/RNAseq-nf GitHub repository. https://github.com/IARCbioinfo/RNAseq-transcript-nf, accessed May 2020.
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014;15(12):550.
40. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30(10):1363–1369.
41. IARCbioinfo/Methylation_analysis_scripts GitHub repository. https://github.com/IARCbioinfo/Methylation_analysis_scripts, accessed July 2019.
42. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology 2016;17(1):208.
43. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv 2018;1802.03426.
44. Konopka T. umap: Uniform Manifold Approximation and Projection; 2019, https://CRAN.R-project.org/package=umap, r package version 0.2.4.0.
45. pan-LNEN TumorMap. https://tumormap.ucsc.edu/?p=RCG_lungNENomics/LNEN, accessed July 2019.
46. Dray S, Dufour AB. The ade4 Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software 2007;22(4):1–20.
47. Martins RM, Minghim R, Telea AC. Explaining Neighborhood Preservation for Multidimensional Projections. In: Borgo R, Turkay C, editors. Computer Graphics and Visual Computing (CGVC) The Eurographics Association; 2015. .
48. Moran PA. Notes on continuous stochastic phenomena. Biometrika 1950;37(1-2):17–23.
49. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2018;35:526–528.
50. TumorMap site. https://tumormap.ucsc.edu, accessed January 2020.
51. IARC bioinformatics platform. https://github.com/IARCbioinfo, accessed January 2020.
52. Nextjournal notebook: A molecular map of lung neuroendocrine neoplasms. https://nextjournal.com/rarecancersgenomics/a-molecular-map-of-lung-neuroendocrine-neoplasms/, accessed January 2020.
53. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, et al. Barriers to accessing public cancer genomic data. Sci Data 2019 06;6(1):98.
54. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, et al. Integrating Genomics into Healthcare: A Global Responsibility. Am J Hum Genet 2019 01;104(1):13–20.
55. Espadoto M, Hirata NST, Telea AC. Deep Learning Multidimensional Projections. arXiv 2019;1902.07958.
56. Supporting data for "A molecular map of lung neuroendocrine neoplasms." GigaScience Database. http://dx.doi.org/10.5524/100781.

Chapter 4

Exploring associations between germline and somatic variations in the lung

4.1 Context

In the past decades, Genome-Wide Association Studies (GWAS) have been performed to identify genetic variants associated with multiple diseases. However, such studies still lack the power to detect small genetic effects, and in most cases, the variants identified so far explain a small proportion of the disease heritability. This issue refers to a notion called the missing heritability [152]. Thanks to the rise of large consortia, the size of GWAS is though being increased, suggesting that the number of susceptibility loci will continue to grow and expand our knowledge on many diseases, including cancer susceptibility. At the same time, the scope of current GWAS study designs and results has been extended [153]. Methods have been developed to meta-analyze GWAS summary statistics to detect novel loci or to investigate genetic correlations [154, 155]. Also, while the efforts focused at first on individual SNPs effects, there is an increasing number of studies attempting to combine the effects of multiple SNPs via the study of epistasis (SNPs interactions) [156] or the use of Polygenic Risk Scores (PRS) analyses [157, 158]. PRS have been recently used to combine the SNPs identified by GWAS in order to explain a larger proportion of the disease risk and identify individual at risks. Indeed, as mentioned previously, GWAS usually reveal SNPs with small effect sizes and complex diseases like cancer are known to be polygenic, *i.e.* multiple genes are involved in the disease development. Another step ahead of the GWAS work consists in revealing the biological mechanisms behind the detected associations [78]. Researchers attempt to identify causal genes using pathways and functional annotations based methods [159], causality models [160] or the integration of GWAS results with other

data resources. The latter analyses essentially focused so far on expression data to perform expression quantitative trait loci (eQTL) and transcriptome-wide association study (TWAS) analyses [161]. eQTL analyses test the impact of SNPs identified by GWAS studies on expression levels in different tissues. TWAS analyses rely on both GWAS and eQTLs results to test associations between gene expression and the disease phenotype to identify causal genes. Those methods have been largely applied on the Genotype-Tissue Expression (GTEx) data, whose 8th version has been recently released [162]. Both methods take advantage of expression data to reveal causal genes but other data types could be considered. Indeed, in parallel to GWAS, exploring the germline susceptibility to cancers, initiatives like the TCGA have generated multi-omics data, improving the molecular characterization of multiple tumor types. Together, these datasets give us the opportunity to explore associations between germline variations and somatic events. Such study design could enable to validate some of the susceptibility SNPs identified by GWAS by providing further support of their causal effects on the disease and bring new insights on the molecular mechanisms involved [163, 164]. So far, multiple studies have identified germline-somatic interactions. Carter *et al.* recently showed that germline events could influence the alteration frequency in cancer related genes [165]. Another example is the identification of germline alterations inducing an enrichment in the APOBEC signature in breast and bladder cancer [166, 167]. In this chapter, we focused on lung cancer susceptibility and its interaction with somatic mutational burden in lung tumors by integrating the results of lung cancer and smoking related traits GWAS with the somatic mutations data from the TCGA lung tumors.

4.2 Research contribution

4.2.1 Introduction

Lung cancer is one of the most common cancer worldwide, with around 2 million new cases in 2018 [23]. About 80% to 85% of lung cancers are non-small cell lung cancers and among them, lung adenocarcinomas (LUAD) and lung squamous cell carcinomas (LUSC) are the most frequent. A major risk factor for both subtypes is smoking. However smoking behaviours differ between the subtypes and around only 15% of smokers develop lung cancers [76] (See Introduction section 1.1), suggesting that smoking is not the only risk factor for this disease. Until now, most research studies aiming at understanding lung cancer etiology and development have focused their efforts either on germline analyses or on somatic analyses separately.

Yet, disentangling the interactions between germline and somatic events could, on one hand, facilitate the identification of the causal genes involved in lung cancer susceptibility and on the other hand bring light on the molecular characteristics of the lung tumors.

In the past decades, multiple GWAS have been performed to identify germline variations associated with lung cancers. The first GWAS were conducted in 2008 and identified a strong susceptibility locus on the chromosome 15q25 region [168, 169, 79] where multiple genes including three nicotinic acetylcholine receptors are located: the *Cholinergic Receptor Nicotinic Alpha 5 Subunit (CHRNA5)*, the *Cholinergic Receptor Nicotinic Alpha 3 Subunit (CHRNA3)* and the *Cholinergic Receptor Nicotinic Beta 4 Subunit (CHRNB4)* genes [78]. The effect of the locus on smoking was assessed and at the time, only one of the three studies suggested a direct effect rather than an indirect effect through smoking [169]. To date, no consensus has been reached with regard to this locus effect [170, 171, 172]. Although the locus on chromosome 15q25 region displayed the strongest effect in each GWAS, two other loci were identified the same year, one on chromosome 6 by Wang *et al.* [173] and one on chromosome 5 by Wang *et al.* [173] and McKay *et al.* [174]. In contrast with the first locus, the latter hits did not associate with smoking and their association with lung cancer differed depending on the lung cancer subtypes considered, LUAD and LUSC [175]. In 2017, McKay *et al.* performed the latest and largest GWAS on lung cancer in European ancestry and in addition to identify new loci, confirmed that lung cancer susceptibility is heterogeneous across histological subtypes [80].

As mentioned before, tobacco smoking is a major lung cancer risk factor, its impact on our cells has been widely studied and is known to cause DNA damages induced by the carcinogens found among the tobacco chemicals [176]. If not repaired, those damages accumulate in the lung tissues, hence increasing the mutational burden of cells and leaving particular mutational patterns in the damaged cells [177]. Such mutational patterns caused by exogenous or endogenous processes are called mutational signatures (See Introduction section 1.1). The most common signature found in smoking related lung cancers is the Signature 4 which has been described to be enriched for $C > A$ substitutions and caused by DNA damages resulting from benzo[a]pyrene exposure, a mutagenic carcinogen found in tobacco smoke [29, 177]. These observations imply that the more a person smokes, the higher the mutational burden in its lung will be, hence the higher its risk of developing lung cancer. This supports epidemiological studies such as those performed by Doll *et al.* showing a 20 fold increase in lung cancer risk in smokers versus non-smokers [25]. While those correlations are known, the direct association between lung cancer germline

variants and somatic mutational burden in lung tumors has, to our knowledge, not been tested. In this work, we investigated the association between lung cancer susceptibility variations and the somatic mutation burden in lung tumors using different approaches (Figure 4.1). Firstly, we built lung cancer genetic risk scores and tested their association with the total number of somatic mutations as well as with the number of mutations attributable to Signature 4, related to smoking exposure. Secondly, the Mendelian randomization setting was used to assess the causal effect between smoking traits and mutational burden in lung tumors.

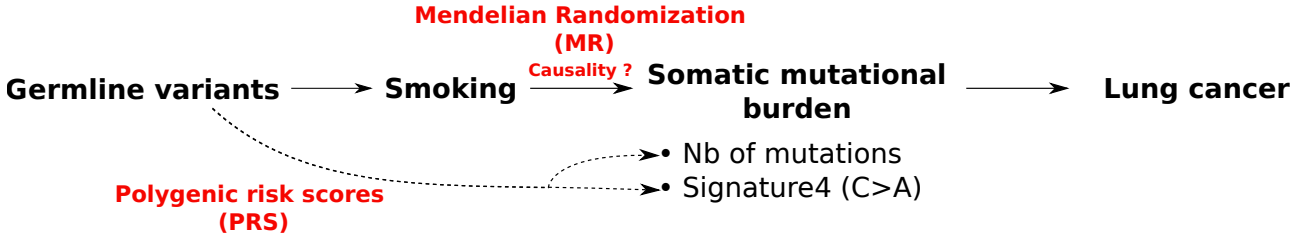


FIGURE 4.1: **Outcome and exposures relationships.** This figure presents the variables involved in the relationship between the exposure, smoking, and the outcome, mutational burden.

4.2.2 Material and methods

The study of the association between germline susceptibility and somatic mutational burden in tumors required to use different sources of data. On one hand, GWAS data were needed to identify variants associated with the traits of interests, *i.e.* smoking traits and lung cancer. On the other hand, samples for which germline and somatic data are available were required to test the effect of the previously mentioned variants on mutation burden. The TCGA dataset met this requirement with multiple omics data available, including genotyping arrays and WES data for 33 cancer types, including lung cancer. For the germline data, the processing and quality control of the raw genotyping arrays data was carried out in order to perform imputation to retrieve the genotyping information at non-assayed positions [178]. In parallel, the public somatic mutations files were processed to derive mutational burden and mutational signature attributions.

Imputation of the TCGA samples

- **Data download and samples selection**
  Genotyping was performed for the samples from the 33 TCGA cohorts on the raw intensities CEL files downloaded on the GDC Legacy Archive portal using gdc-client (v1.4.0) [179]. Only blood and/or normal tissue samples with DNA

analyte were downloaded (in total 11837 files). The CEL files were then filtered out if associated with one of the following TCGA non-rescinded annotations: "Item flagged DNU", "Administrative Compliance", "Item does/may not meet study protocol", "Qualified in error", "BCR Notification", "Normal tissue origin incorrect", "Subject withdrew consent", "Normal class but appears diseased", "Duplicate item", "Tumor tissue origin incorrect", "Tumor type incorrect", "Genotype mismatch", "Permanently missing item or object" and "Qualification metrics changed". Also, when multiple samples were available for a participant, blood samples were selected when available. 10443 samples remained after filtering on annotations and removing duplicated samples.

- **Genotyping**
  Prior to genotyping, a quality control function, *apt-geno-qc*, from Affymetrix Power Tools (APT) has been applied to each cohort separately. As recommended by Andrade *et al.* [180], samples with a contrast QC (CQC) value of at least 0.4 were kept for genotyping. The largest TCGA cohort, Breast Invasive Carcinoma (BRCA), was then considered to define a list of probes with good genotyping call rate (above 97%), used subsequently to genotype each sample using the *apt-probeset-genotype* function and the birdseed algorithm. Each TCGA cohort was genotyped a first time and all samples with genotyping call rate lower than 97% were excluded. After exclusion of the samples with low genotyping quality, a second round of genotyping was performed on each cohort. The genotyping outputs were converted to the plink format using *apt-result-format* from the APT tools (version 2.10.2.2) with the following annotation "GenomeWideSNP_6 .na35.annot.db" provided by the Affymetrix Support by Product web page [181]. The sex column included in the pedigree file was retrieved using the curated TCGA clinical data from Liu *et al.* study [182]. Finally, the plink files associated with each TCGA cohort were merged in a single dataset using the *merge-list* option from plink (v1.90b4).

- **Origin inference**
  After controlling for the plink files integrity using the *HRC-1000G-check-bim.pl* script available on the McCarthy web site (version 4.2.11) [183], the origin of each sample was predicted using admixture [184] (version 1.3) in a supervised mode. The HapMap Phase II dataset was considered as a reference dataset [185] and the list of SNPs defined by Yu *et al.* in 2008 was used for the origin inference [186], among 12898 SNPs, 11630 were in common between the HapMap and the TCGA datasets. The number of origins to infer (K) was

fixed to three (Europeans, Africans, Asians). Admixture thus assigned, to each sample, probabilities of belonging to each of the three population group. The group with the highest probability defined the samples inferred ancestry. Around 99% of the samples reported as "WHITE" by the TCGA clinical data were predicted Europeans, around 90% of the samples reported as "ASIAN" were predicted Asians and 95% of the samples reported "BLACK OR AFRICAN AMERICAN" were predicted Africans.

- **Samples statistics**
  In each ancestry group, reported sex for each sample was compared to imputed sex based on genotyping data using the *check-sex* option from plink (v1.90b4). Samples relatedness was tested using the *genome* option from plink (min parameter fixed at 0.185). Among the 9855 samples with reported sex, 46 samples with discordant reported sex and imputed sex and 17 pairs of relatives were identified and flagged. Finally, plink was also used to compute heterozygosity and genotyping missing rates (*het* and *missing* options, respectively). Figure 4.2 represents, for each origin, the concordance between reported and predicted sex across samples as well as the heterozygosity rate as a function of the missing rate and shows that in each ancestry group, all samples had a genotyping missing rate lower than 3% and a homogeneous heterozygosity rate.

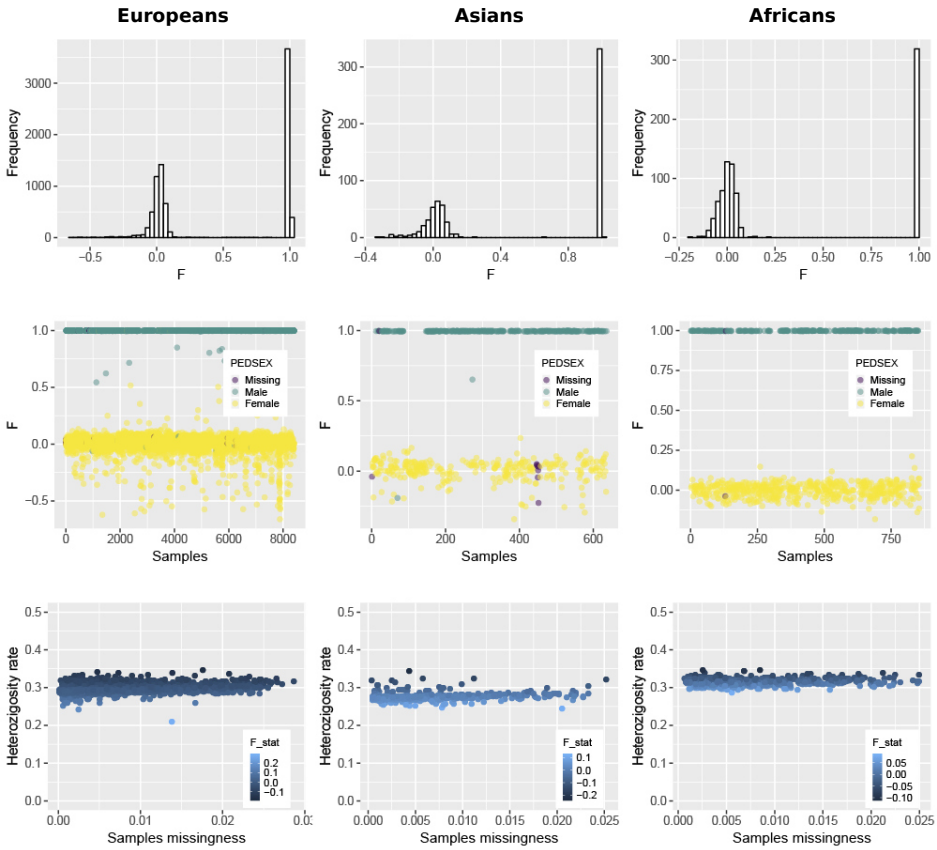


FIGURE 4.2: **Samples quality control.** For each ancestry group, the sex checking F statistic distribution across samples is represented in the upper panel ($F < 0.2$ and $F > 0.8$ leading to females and males calls respectively), the concordance between the reported sex (points colors) and the statistic is shown in the middle panel, and the heterozygosity rate as a function of the missing rate in the bottom panel.

- **SNPs filtering**

  SNPs were filtered out based on different criteria. Firstly, SNPs with genotyping call rate lower than 97% across all the TCGA samples and with Minor Allele Frequency (MAF) below 1% were excluded using plink. Secondly, since SNPs allele frequencies vary depending on samples ancestry, the TCGA dataset was split into different groups based on the origin inferred by admixture. In each ancestry group, we applied the *HRC-1000G-check-bim.pl* script from the McCarthy tools [183], which allows to remove the SNPs with unmatched positions and/or alleles, duplicated SNPs, ambiguous SNPs with MAF above 40% as well as SNPs with allele frequencies differing from the allele frequency reported in the same population in the 1000 Genome dataset (difference of more than 20%). SNPs with a genotyping call rate below 97%

and showing strong deviation (p-value < 10-8) from the Hardy Weinberg equilibrium (*hwe* plink option) in any of the ancestry groups were excluded. Finally, ambiguous SNPs were not considered. The Figure 4.3 represents for each ancestry group the Alternative frequency (AF) of the remaining SNPs in the TCGA dataset as a function of the SNPs AF in the 1000 Genome dataset in the same population (left panels) and shows a high correlation between the two datasets.

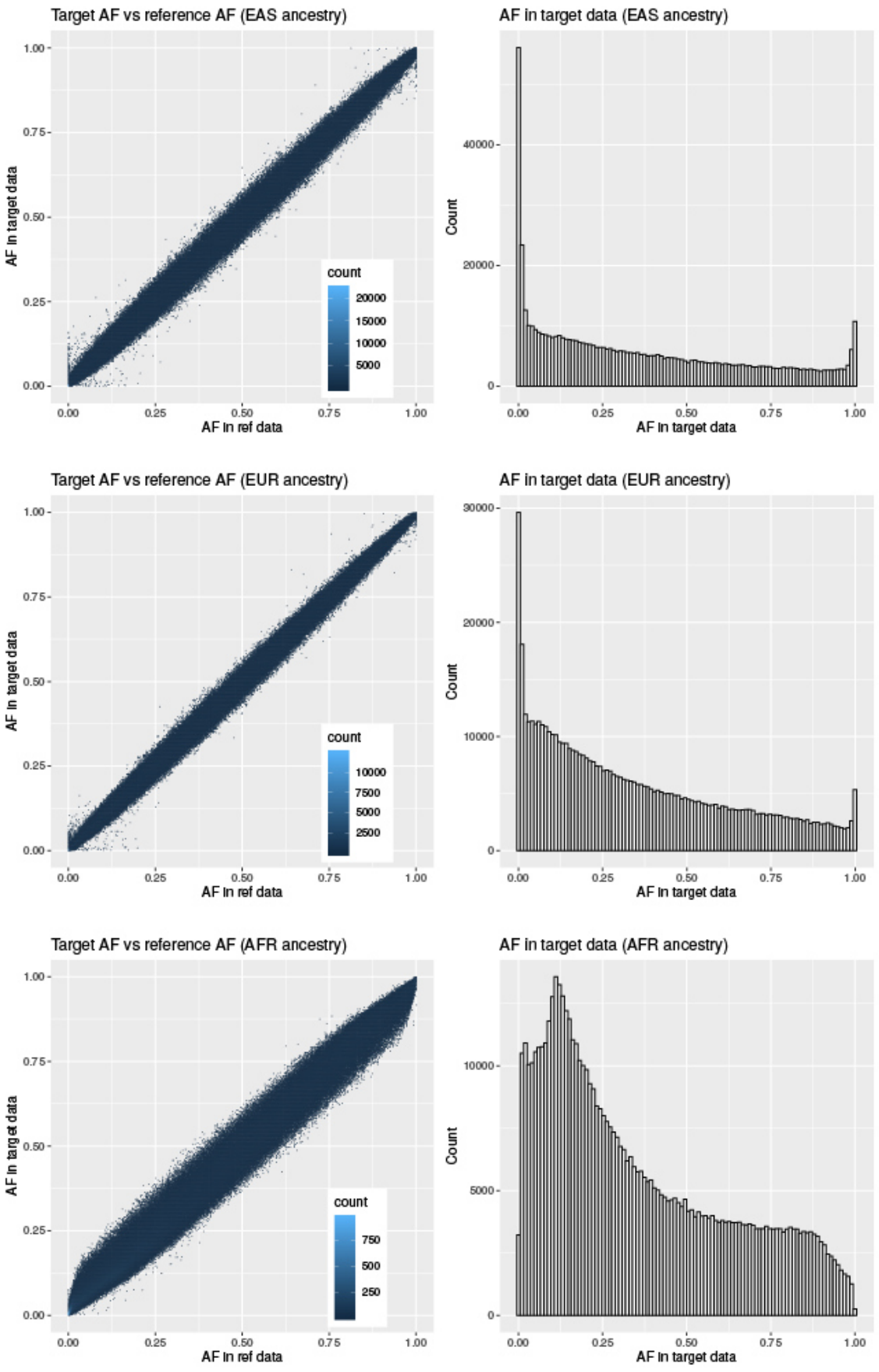


FIGURE 4.3: **SNPs filtering quality control.** For each ancestry group (EAS: Asians, EUR: Europeans, AFR: Africans), the left panel represents the AF of the remaining SNPs, after filtering, in the TCGA dataset, as a function of the SNPs AF in the 1000 Genome dataset. The right panel corresponds to the distribution of AF of all SNPs.

- **Imputation**
  After SNPs filtering, phasing and imputation were performed on each chromosome. Phasing was performed using eagle (v2.4.1) [187] and the 1000 Genome phase 3 data as reference; the reference is available on the International Genome Sample Resource (IGSR) as VCF files [188]. Bcftools (v1.8) was used to convert the VCF files to Binary Variant Call Format, (BCF) files, select SNPs and indels and to normalize variants in order to consider multi-allelic positions. Eagle was run on each chromosome divided by chunks of 20 Mb using a flanking region of 5 Mb. The resulting phased VCF files served as input to minimac4 (v1.0.1) [189], which performed the imputation with a window of 500 kb. Figure 4.4 provides a graphical representation, in the European ancestry group, of the imputation quality. In the left panel, the distribution of minimac4 R2 quality measure is represented for three categories of SNPs: SNPs with MAF > 5%, SNPs with MAF between 0.5 and 5% and SNPs with MAF below 0.5%. We also compared the allele frequencies of the SNPs (with an R2 value above 0.3) in the imputed data, with the allele frequencies of the same SNPs in the 1000 Genome dataset (right panel) and observed that they were correlated. The Annex Figure C.1 to C.3 complete the previously mentioned figures with all ancestry groups in the TCGA data.

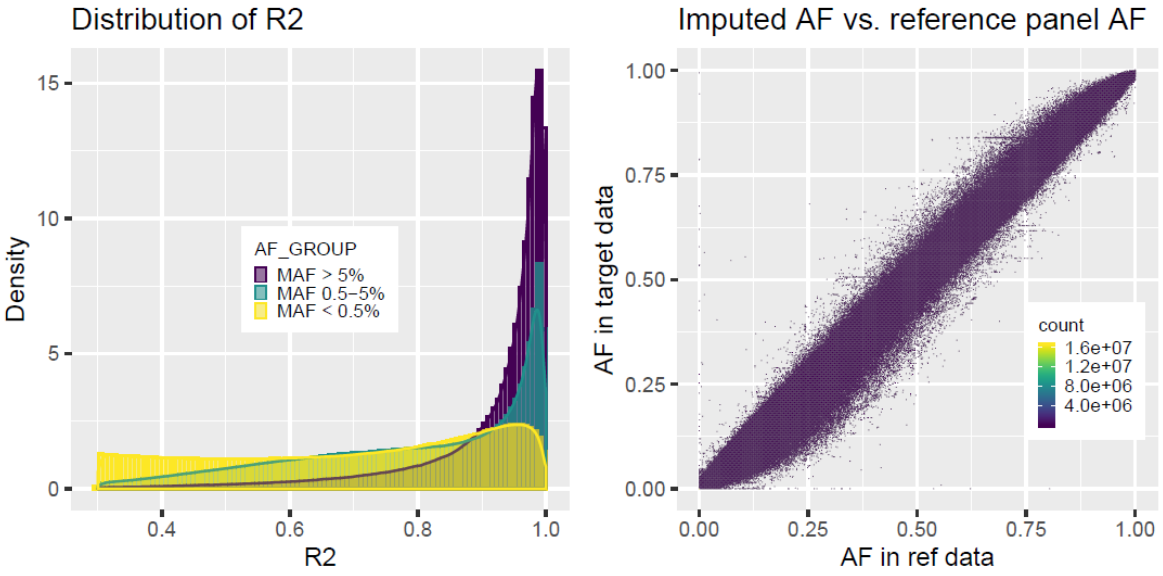


FIGURE 4.4: **Imputation quality controls (European samples)**. Left panel: minimac4 R2 quality measure distribution for each MAF category (MAF > 5%, MAF between 0.5 and 5% and MAF below 0.5%). Right panel: comparison of the SNPs AF (with an R2 value above 0.3) in the imputed data with the same SNPs AF in the 1000 Genome dataset.

Finally, since population stratification can be a confounding factor in genetic studies, it is necessary to include ancestry variables as covariates in regression models. The main ancestry population in the TCGA cohorts is the European population. We thus selected the samples predicted as being Europeans by admixture and ran the software Eigenstrat [190] to correct for population structure among Europeans. We considered the list of SNPs defined by Yu *et al.* in 2008 [186] to run Eigenstrat without outliers removal.

Imputation pipeline

While the work presented in this study focused on the germline somatic interactions in lung cancer, the data processing has been performed on all 33 TCGA cohorts (more than 9000 samples). The code used to perform the imputation of those samples has been adapted by a master student, which I co-supervised during five months, in an automatized, reproducible and portable nextflow pipeline that is publicly available on the IARCbioinfo/Imputation-nf GitHub repository. A docker and a singularity container have been generated to allow future users to run the pipeline without having to install the needed softwares. The pipeline performs the quality controls and data processing necessary to carry out imputation locally but also to submit imputation jobs to the Michigan imputation server [191] as well as to the recently developed TopMed imputation server [192]. Finally, quality control figures similar to those presented previously are generated by the pipeline automatically (See Figures 4.2, 4.3 and 4.4).

Figure 4.5 provides an overview of the study design and analyses performed based on the imputed data obtained following the steps described in the previous paragraphs.

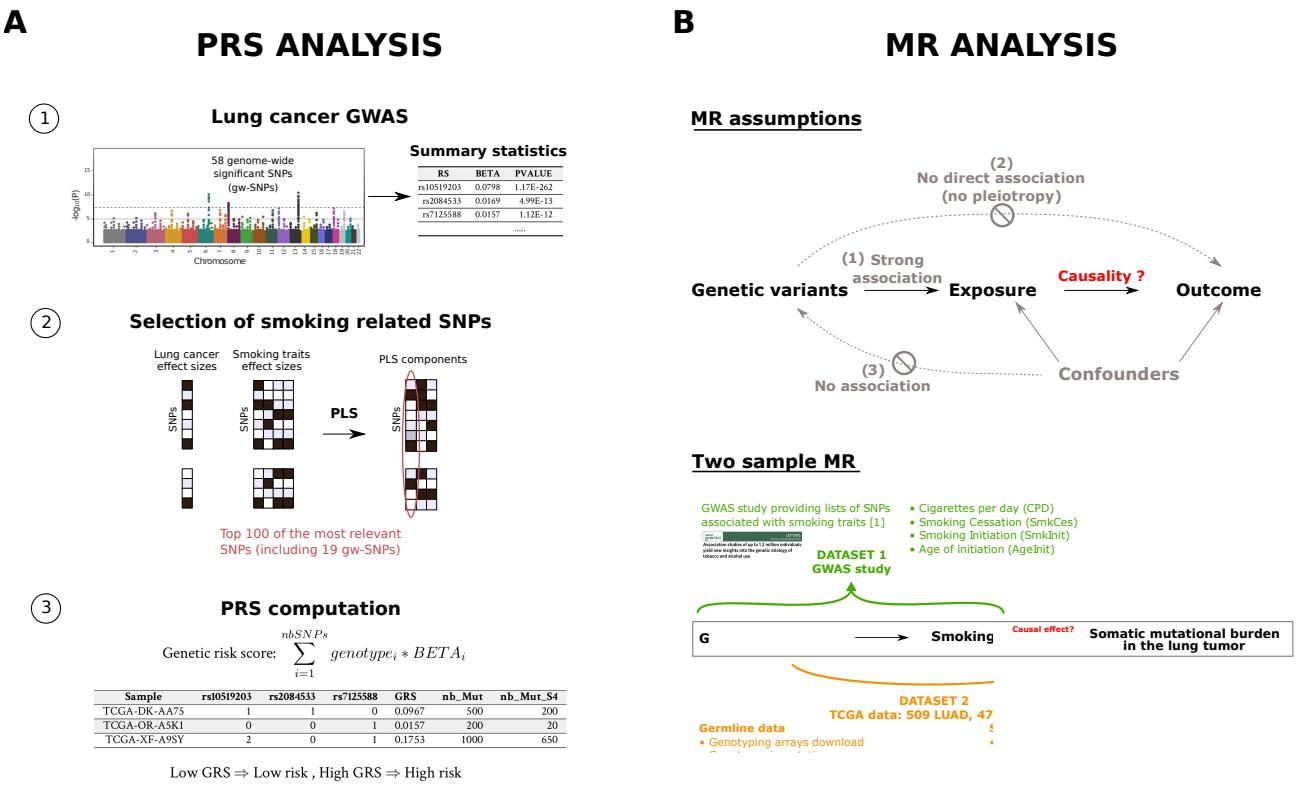


FIGURE 4.5: **Methods overview and study design.** A) Steps of the PRS analysis. A meta-analysis of lung cancer GWAS was performed (1). The results of the meta-analysis were used in conjunction with smoking traits GWAS to select different lists of SNPs to compute PRS (2 and 3). The association of these scores with mutational burden were then tested. B) Mendelian Randomization (MR) analysis assumptions and design (top and bottom panel, respectively). A two sample MR setting was used to test the causal effect of smoking exposures on mutational burden in lung tumors.

Lung cancer GWAS data

In order to increase the sample size of existing lung cancer GWAS, a family history GWAS (GWAx) expanding traditional GWAS to familial cases, was performed on the UKbiobank data. The resulting analysis was meta-analyzed with a previously published GWAS on lung cancer based on the Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) dataset [80], the meta-analysis was performed using the software *metasoft* with fixed-effects model [193]. The SNPs resulting from the meta-analysis were pruned using plink to remove SNPs in linkage disequilibrium (r^2 threshold fixed at 0.1) and used for the selection of lung cancer related SNPs. While the meta-analysis revealed multiple significant lung cancer-associated hits (Figure 4.5A step 1), we attempted to select relevant SNPs that did not pass the genome-wide significance threshold for further

PRS analyses. Indeed, while PRS usually aggregate the effects of the genome-wide significant SNPs, previous studies have shown that additional information could be retrieved from other variants [194, 195]. In this study, we chose to select additional SNPs based on their association with smoking related traits. For that purpose, summary statistics of previously published GWAS on smoking related traits have been gathered (Figure 4.5A step 2). We chose the traits studied by the GWAS Sequencing Consortium of Alcohol and Nicotine (GSCAN) consortium in a large dataset gathering up to 1.2 million samples [196]. This study explored four smoking traits (cigarettes per day, smoking cessation, smoking initiation and age of initiation) as well as drinking consumption. It has been shown by Jiang *et al.* that there is a shared heritability, probably mostly driven by smoking, between lung cancer and head and neck cancer [197]. Therefore, the summary statistics of the GWAS on head and neck cancer performed by Lesseur *et al.* was also considered [198]. Our hypothesis was that a SNP associated with one or more of those traits and associated with lung cancer could be valuable in a PRS predicting lung cancer. In order to select such SNPs, the partial least square (PLS) model, which can be assimilated to a supervised PCA (See introduction section 1.4), was used considering the Z scores (associations effect sizes divided by the standard error) for each trait as explanatory variables and the lung cancer Z scores as the response variable. Hence, this method generated latent components that maximize the covariance between the smoking related traits summary statistics and the lung cancer summary statistics. The first component, positively correlated with the smoking trait, was used to rank the lung cancer GWAS SNPs. The top 100 SNPs with the highest values on this component were selected as the most relevant SNPs, the later list of SNPs will be referred to as the smoking related SNPs. In addition to the genome-wide significant SNPs, the smoking related SNPs were considered to compute and test the association of different PRS scores with the mutational burden in the TCGA samples (See next paragraph on PRS computation).

PRS computation and regression analyses

We built PRS using different lists based on the aforementioned SNPs using PRSice2 [199], which computes PRS values as a weighted sum of the SNPs imputed dosages, the weights being the effect size of the SNPs on lung cancer (Figure 4.5A step 3). Firstly, three PRS were generated based on the lung cancer genome-wide significant SNPs (gw-SNPs): i) a PRS based on all the gw-SNPs, ii) a PRS based on the smoking related gw-SNPs (19 out of the lung cancer gw-SNPs ranked in the top 100 smoking related SNPs), iii) a PRS based on the non-smoking gw-SNPs. Secondly, three PRS

were computed based on the top 100 smoking related SNPs described in the previous paragraph: i) a PRS including the 100 SNPs, ii) a PRS based on the gw-SNPs in that list, iii) a PRS based the non gw-SNPs in that list. In each case, we tested the PRS association with the total number of mutations in the tumors as well as with the number of mutations attributable to Signature 4, which is related to smoking. Based on the skewed distribution of the two mutational burden variables (See Figure 4.6), the negative binomial regression, usually used for over-dispersed count variables, was chosen to test the association (*glm.nb* function from the R package MASS). Multiple covariates were included in the model: age, gender and the 10 first principal components resulting from Eigenstrat. The samples purity provided by the Pan-Cancer Atlas files [52], was transformed in a categorical variable (purity less than or equal to 30%, purity between 30 and 70% included and purity above 70%) and was added to the covariates as well. Finally, when the LUAD and LUSC cohorts were both considered, a categorical variable indicating the sample's cohort was included in the model.

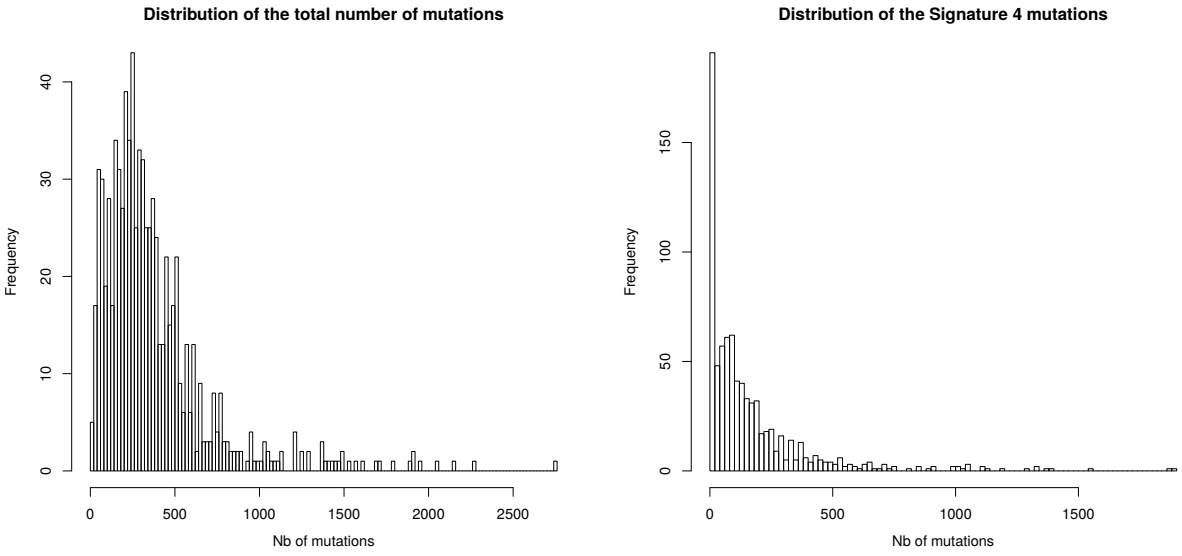


FIGURE 4.6: **Distribution of the mutational burden in the TCGA lung cancer samples**. Distribution of the total number of mutations (left panel) and the number of signatures attributable to Signature 4 (right panel) in the TCGA lung cancer samples.

Mutational burden and mutational signatures

The somatic mutations of the TCGA samples were retrieved from the study of Ellrott *et al.* [200] performed in the context of the Multi-Center Mutation Calling in Multiple Cancers (MC3) project. The MC3 Mutation Annotation Format (maf) files gather the curated results from seven different variant callers for all TCGA samples and report

mutations found by at least two variant callers. In order to remove duplicated samples per patient, the variants flagged "nonpreferedpair" were removed (10224 samples left). Samples who were whole genome amplified or flagged "gapfiller"were excluded from the analyses. The total number of mutations for each sample was computed based on the set of filtered variants. Finally, the proportion of artifacts or germline variants were computed to identify and remove potential low quality samples (proportion higher than 10%) [201].

The filtered maf file was split into one maf file per cohort and converted to the VCF format using a perl program downloaded from the following GitHub repository: mskcc/vcf2maf. Based on those VCF files, signature contributions have been computed using MutationalPatterns [202]. The software computes the contributions of the mutations to the known COSMIC signatures version 2 (30 signatures) based on non-negative least squares (NNLS) method. The number of mutations attributable to a signature has been computed by multiplying the total number of mutations by the related signature contribution.

RNA-Seq data

Expression data were firstly used to identify potential misclassified samples that could bias the association tests when stratifying the lung samples by histological subtypes. For that purpose, the dimensionality reduction method UMAP [111] was run on both TCGA lung cancer cohorts. The raw RNA-Seq data were processed from the alignments of the reads to the reads counts computation using in-house Nextflow [141] pipelines available at the following GitHub repositories: IARCbioinfo/RNAseq-nf release v2.3, IARCbioinfo/abra-nf release v3.0, IARCbioinfo/BQSR-nf release v1.1 and IARCbioinfo/RNAseq-transcript-nf release v2.1. The read counts were normalized using the variance stabilization transformation (*vst* function from DESeq2 R package), sex and mitochondria chromosomes were removed, and the most variable genes, explaining 50% of the variance were kept to run UMAP. The UMAP representation of the samples allowed us to observe two distinct clusters representing the two lung histopathological subtypes (See Annex Figure C.4). Samples with unexpected molecular clustering (LUAD samples clustering with LUSC samples and vice versa) were excluded from the regression analysis.

The RNA-Seq data were also used to perform Gene Set Variation Analysis (GSVA) in order to identify differential pathways activities between carriers and non-carriers

(homozygote groups) of the rs10519203 SNPs located on the chromosome 15q25 region. The analysis was run using the R package GSVA [203]. Ten genes sets describing the hallmarks of cancer [204] were considered. The *gsva* function computed for each sample and each gene set enrichment scores, which measures the enrichment of the genes inside a gene set in comparison to the genes outside the gene set, and using the R package limma [205], the difference between the GSVA enrichment scores of the two homozygote groups was tested.

Mendelian randomization (MR) analyses

PRS computation was used to test the association between lung cancer susceptibility and mutation burden. This method however does not enable to test the causal mechanisms underlying the observed associations. For that purpose, we used MR methods that attempt to test a causal effect between an exposure and an outcome using genetic variants, also called genetic instruments, as proxy for the exposure [206, 207, 208]. The idea is that, if genetic variants are associated with an exposure that is causal for a disease, the same SNPs should be associated with the disease. The MR analysis setting could be compared to Randomized Control Trials (RCT) with groups being formed based on genetics. Indeed, following Mendel's second law, we can assume that the genetic variants are allocated at birth randomly. Thus, they should not change over time due to environmental exposures, and confounders issues faced in classical RCTs should hence be less problematic. However, three main assumptions need to be respected [208] (Figure 4.5B top panel): i) the relevance assumption, which implies that the genetic instruments do strongly associate with the exposure, ii) the independence assumption, which implies no association of the genetic variants with confounders of the exposure-outcome relation and iii) the exclusion restriction assumption, which implies that the genetic variants do not impact the outcome variable via other pathways than the exposure pathway (pleiotropic effect).

In order to assess the causal link between smoking traits and mutational burden, a two sample MR analysis was performed (Figure 4.5B bottom panel). The two sample MR setting estimates the effects of the genetic variants on the exposure and on the outcome in two independent datasets. For the dataset intended to select the genetic variants associated to the smoking traits, we chose the GSCAN study which is based on up to 1.2 million samples [196] and retrieved the variants strongly associated (p-value above 10-8) with cigarettes per day (CPD), smoking cessation (SmkCes), smoking initiation (SmkInit) and age of initiation (AgeInit). In this study, only European ancestry samples were considered. The second dataset used

to determine the association between each selected variant and the mutational burden was the TCGA dataset, for which both germline and somatic data (described in the previous paragraphs) are available.

In order to test a causal effect using MR methods, the use of multiple MR tests relying on different assumptions is recommended [208] to ensure that the results are reliable. We used the TwoSampleMR R package [209] to perform five tests (Inverse Variance Weighted (IVW), weighted median, weighted mode and MR-Egger tests). The MR-Egger test allowed to test also if pleiotropic effects are involved in the studied associations. Finally, to assess if the associations observed were driven by a few SNPs, a leave-one-out analysis based on the IVW method was performed.

4.2.3 Results

Family history GWAS

The meta-analysis of the family history GWAS and the TRICL-ILCCO lung cancer GWAS identified 65 genome-wide significant hits associated with lung cancer, those hits being located in 20 distinct genomic regions. Among the genome-wide significant (gw-SNPs), multiple SNPs were also associated to smoking traits, like the number of cigarettes smoked per day, smoking cessation and initiation, while others were related to other pathways like DNA repair. Considering these observations, different PRS scores were built and their association with mutational burden were tested. For that purpose, the gw-SNPs were stratified in smoking and non-smoking related SNPs. For the PRS based on all the gw-SNPs, no association with mutational burden was observed (Figure 4.7, $\beta = 0.03$ and *p-value* $= 0.373$). However, when including only the smoking related SNPs in the PRS computation, associations with the total number of mutations as well as with the number of mutations attributable to Signature 4 were observed (Figure 4.7, β of 0.09 and 0.13 and *p-value* of 0.002 and 0.026 respectively). In order to determine if the significant associations observed were driven by a restricted number of SNPs, a leave-one-out analysis was performed on the previously mentioned list. This analysis revealed that removing one SNP (rs72740955) located near the *CHRNA5* gene, a nicotinic acetylcholine receptor subunit on the chromosome 15q25 region, dissolved the association between the PRS and the mutational burden (*p-value* $= 0.278$ for the association with the total number of mutations, and *p-value* $= 0.505$ for the association with the Signature 4 mutations). These results suggested that the association observed was driven by

this variant and that combining the effects of several SNPs for this PRS did not improve the predictive power. We then assessed if increasing the number of SNPs by including additional SNPs that did not reach genome-wide significance could add valuable information to the PRS. Using additional summary statistics from GWAS on smoking related traits (See method section), the selection method resulted in a list of 100 relevant SNPs (smoking related SNPs). This list showed the strongest association with mutational burden and Signature 4 mutations (Figure 4.7, β of 0.14 and 0.17 and *p-value* less than 0.0001 and 0.003 respectively). Also, leave-one-out analysis performed on this list showed that the combination of several SNPs, in this case, was valuable. Indeed, the removal of each individual SNPs did not dissolve the observed associations with the total number of mutations nor with Signature 4 mutations (all *p-values* bellow 0.001 and 0.027 respectively), including the removal of the chromosome 15 locus, identified previously as driver.

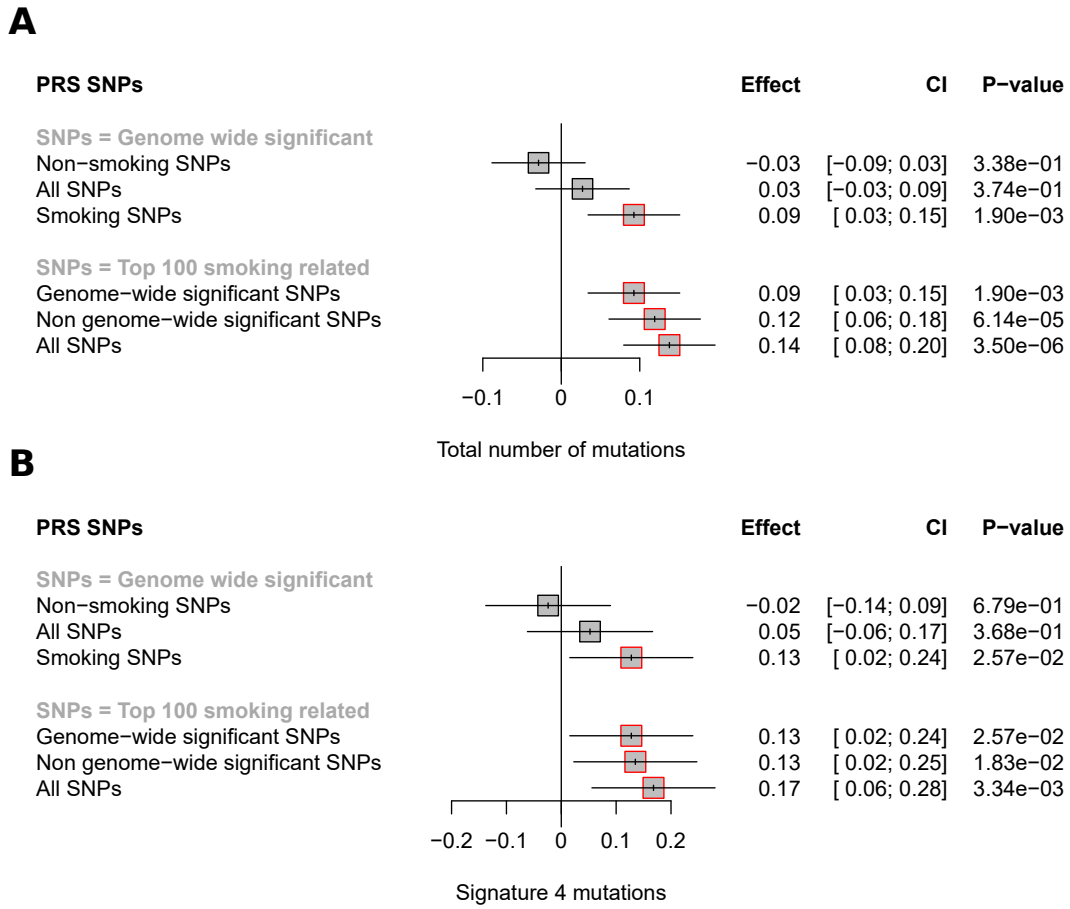


FIGURE 4.7: **Forest plots representing the associations between the PRS scores and mutational burden**. Results of the associations between the PRS scores and the total number of mutations are represented in panel A and with the number of signatures attributable to Signature 4 in panel B.

Differences between LUAD and LUSC samples

To assess if the association observed in the lung cancer samples was consistent in LUAD and LUSC samples separately, we split the lung cancer samples into two groups based on their histological subtype. Figure 4.8 highlights differences observed in LUAD and LUSC samples. Firstly, Figures 4.8A and B show that the PRS built on the smoking related SNPs is associated with the total number of mutations as well as with the Signature 4 mutations only in the LUAD samples. Figure 4.8C represents the distributions of the number of mutations in the different smoking groups (never, current, former smokers who quit smoking since more than 15 years and former smokers who quit smoking since less than 15 years) in each subtype. The mutational burden was more variable between smoking categories in LUAD samples than in LUSC samples. There were though fewer never smokers in the LUSC cohort in comparison to the LUAD cohort (2% of never smokers in LUSC versus 14% in LUAD), which could explain the differences observed between the two cohorts. We thus tested the association between the PRS and mutational burden when stratifying by smoking status (Figures 4.8A and B). In ever smokers, an association between the PRS values and mutational burden was still observed. Nevertheless, the strength of the association decreased and differences of associations were observed when further stratifying in current and former smokers. Indeed, no association between the PRS and mutation burden was observed in the current smokers.

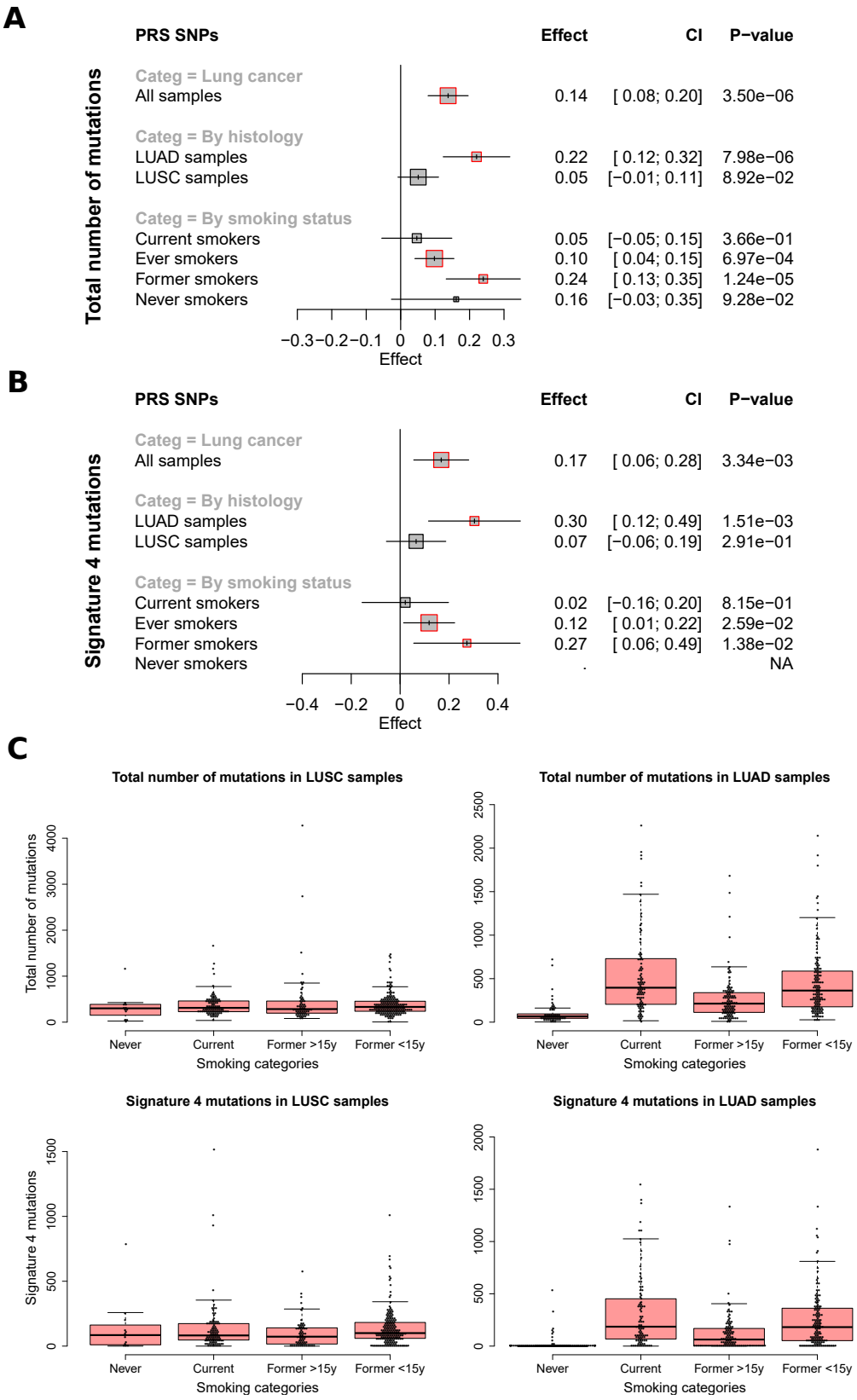


FIGURE 4.8: **Comparison of LUAD and LUSC samples**. Panels A and B show respectively the forest plots representing the associations of the smoking PRS with the total number of mutations and the number of mutations attributable to Signature 4 when stratifying by histology and smoking status. Panel C represents the distributions of the total number of mutations (top) and Signature 4 mutations (bottom) in the two cohorts across different smoking categories (never, current, former smokers who quit smoking since more than 15 years and former smokers who quit smoking since less than 15 years).

Mendelian randomization (MR) analysis

The previous results suggested that the association between lung cancer susceptibility SNPs and mutational burden is strongly related to smoking. Therefore, we selected smoking instruments previously identified for cigarettes per day (CPD), smoking cessation (SmkCes), smoking initiation (SmkInit) and age of initiation (AgeInit) by the GSCAN consortium [210] to validate the causal effect of smoking on mutational burden and to test for potential pleiotropic effects. Multiple MR tests were performed to test the causal effects of each smoking trait on mutational burden. Figure 4.9 provides a graphical representation of those tests results by representing the effects of the genetic instruments on the mutational burden as a function of their effects on each smoking exposure variable. Table 4.1 provides the associated summary statistics. Among the four exposures, only the CPD trait was attributed a significant causal effect on mutational burden in at least two different MR tests. Indeed, out of the five tests performed four concluded on a causal effect. One of these tests was the MR-Egger test that, on top of confirming the causal effect detected by the other tests, identified a pleiotropic effect (Intercept *p-value* of 0.0328). Also, on the scatter plot related to the CPD exposure represented in Figure 4.9, one can notice a potential outlier SNP, with a higher effect on the exposure, that could drive the association. The outlier SNP, rs10519203, is located on the chromosome 15q25 locus identified as driver in the PRS analysis. Following this observation, a leave-one-out analysis was performed and confirmed that removing this variant dissolves the causal effect (IVW test, $p\text{-}value = 0.86$). The heterogeneity observed between the genetic variants coupled with the MR-Egger test results suggests a pleiotropic effect involved in the association between smoking and mutational burden and attests the complexity of the smoking trait.

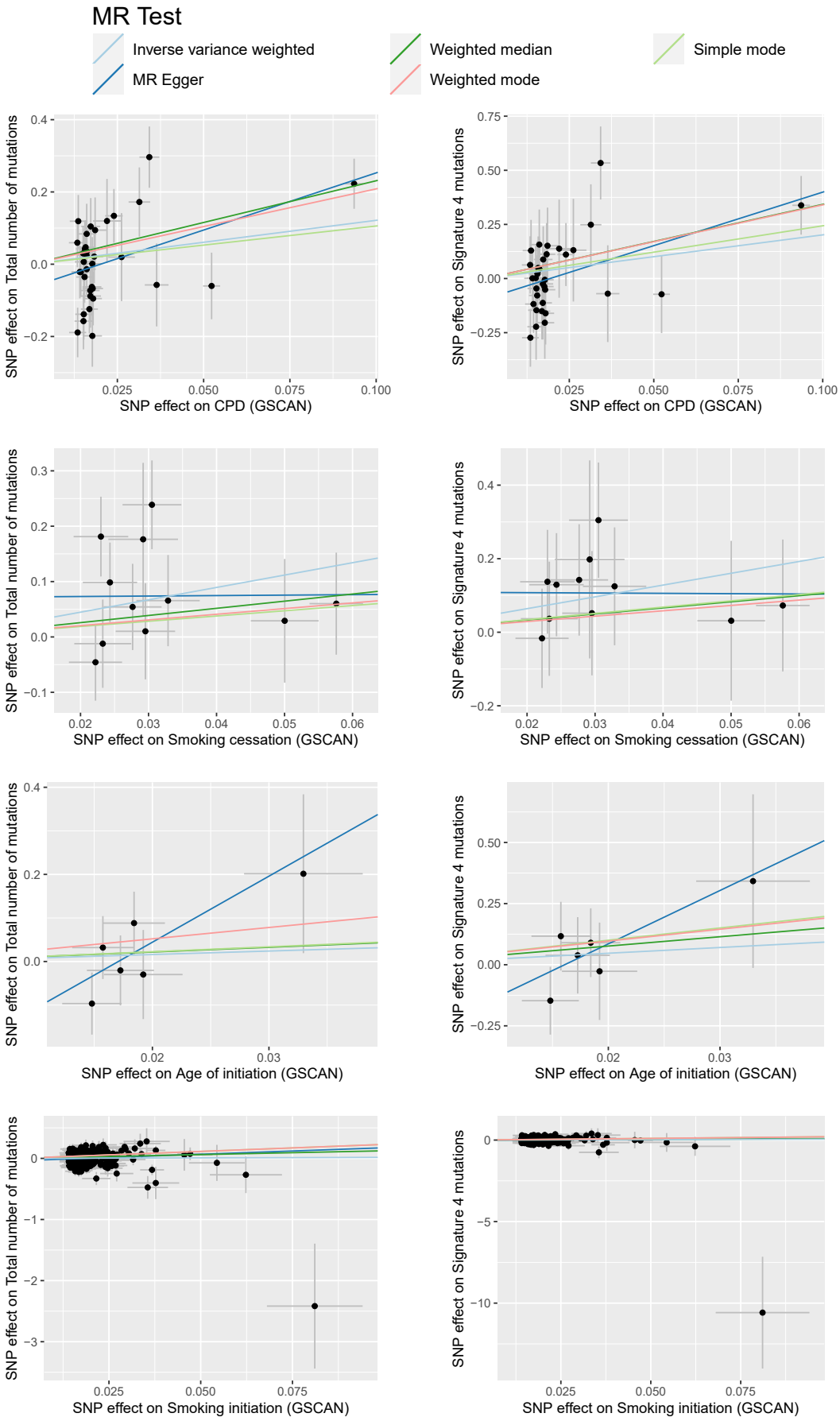


FIGURE 4.9: **Graphical representation of the MR tests results.** Scatter plots representing the effects on the total number of mutations and Signature 4 mutations on the y-axes (left and right columns respectively) and the effects on the four exposures on the x-axes. The colored lines correspond to the regression lines of the different MR tests.

			Total number of mutations			Signature 4 mutations		
Exposure	**Method**	**NbSNP**	**b**	**se**	**pval**	**b**	**se**	**pval**
CPD	MR Egger	35	3.16	1.09	0.007	4.95	1.65	0.005
CPD	Weighted median	35	2.31	0.72	0.001	3.43	1.39	0.014
CPD	IVW	35	1.21	0.69	0.077	2.02	0.99	0.041
CPD	Simple mode	35	1.06	2.15	0.626	2.44	3.36	0.474
CPD	Weighted mode	35	2.09	0.73	0.007	3.4	1.24	0.01
Smoking cessation	MR Egger	11	0.08	2.9	0.978	-0.09	4.88	0.986
Smoking cessation	Weighted median	11	1.29	1.03	0.21	1.66	1.95	0.396
Smoking cessation	IVW	11	2.23	0.9	0.013	3.21	1.55	0.038
Smoking cessation	Simple mode	11	0.94	1.52	0.547	1.71	3.03	0.585
Smoking cessation	Weighted mode	11	1.02	1.32	0.455	1.46	2.59	0.586
Age of initiation	MR Egger	6	15.16	10.18	0.211	21.83	19.82	0.333
Age of initiation	Weighted median	6	1.09	2.51	0.664	3.82	4.59	0.405
Age of initiation	IVW	6	0.8	1.92	0.677	2.35	3.75	0.531
Age of initiation	Simple mode	6	1.12	4.04	0.792	5.01	6.49	0.475
Age of initiation	Weighted mode	6	2.6	4.04	0.548	4.84	5.95	0.453
Smoking initiation	MR Egger	228	2.13	1.35	0.116	1.05	2.17	0.631
Smoking initiation	Weighted median	228	1.24	0.43	0.004	0.81	0.79	0.305
Smoking initiation	IVW	228	0.17	0.33	0.61	0.56	0.53	0.289
Smoking initiation	Simple mode	228	2.29	1.36	0.094	2.06	2.37	0.387
Smoking initiation	Weighted mode	228	2.29	1.14	0.046	2.06	1.91	0.283

TABLE 4.1: **MR summary statistics**. MR summary statistics associated to five MR tests assessing the causal effects of four smoking traits on total number of mutations and Signature 4 mutations.

Both the PRS and MR analyses highlighted the locus on chromosome 15q25 as driver of the associations observed between lung cancer germline susceptibility and somatic mutation load. Figure 4.10 represents the distribution of the total number of mutations (panel A) and the Signature 4 mutations (panel B) in the different genotype group for the rs10519203 SNP in LUAD samples and shows a substantial increase in both variables respectively between the two homozygote groups. Further analyses would be required to understand how these extreme groups differ on the molecular level.

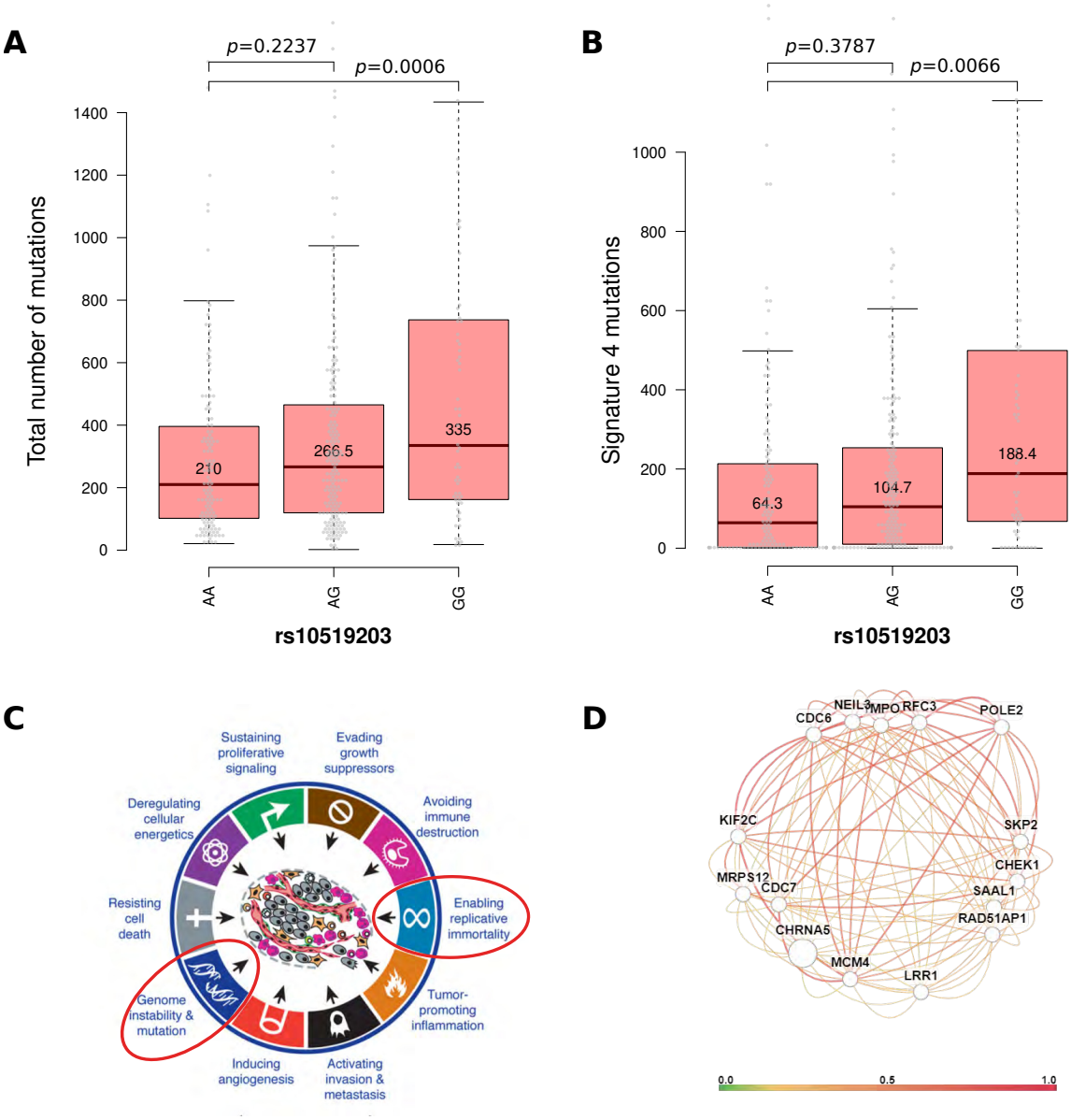


FIGURE 4.10: **The chromosome 15q25 region.** A) Distribution of the total number of mutations in the three genotype groups for rs10519203. B) Distribution of the number of Signature 4 mutations in the three genotype groups for rs10519203. C) The ten hallmarks of cancer, red circles highlight the hallmarks differentially expressed between the rs10519203 homozygote groups (*q-value* threshold set at 0.1). D) Tissue-specific functional networks retrieved from the HumanBase web site [211] representing the genes functionally related to the *CHRNA5* gene in the lung tissue.

4.2.4 Conclusion and discussion

In the past years, multiple GWAS studies have identified more than 40 lung cancer susceptibility loci [78]. The impact of germline variations on lung tissue in conjunction with smoking still needs though to be explored. It is expected that an increase in smoking increases as well the number of somatic mutations in the tumor and thus induces a higher risk of lung cancer. In this study, we confirmed this hypothesis by observing a significant correlation between the number of somatic mutations and lung cancer susceptibility variants. However, the associations observed were heterogeneous across variants and different between the LUAD and LUSC subtypes which suggests that other mechanisms might be involved.

The PRS analysis indicated an association between lung cancer germline susceptibility and somatic mutations in lung tumors. However, this association was dependent on the SNPs included in the PRS. Among the genome-wide significant SNPs, one variant on the chromosome 15q25 locus had a particularly strong effect and drove the germline-somatic interaction in the lung tumors. This observation is in line with multiple previous GWAS analyses highlighting the same region containing the hits with the strongest effects on lung cancer and smoking traits [168, 169, 79, 80]. Nevertheless, the selection of SNPs above the genome-wide significance threshold did add independent information, as the association remained after exclusion of individual SNPs in the leave-one-out analysis. This observation suggested that a better selection of the SNPs could be valuable to explore germline-somatic interactions. One solution to make this selection is to use prior information on each SNP based on other datasets. In this study, we chose GWAS summary statistics of smoking related traits but other data types, like eQTL data or SNPs annotations, *e.g.* the SNPs impact or their associated genes, could be used. A lung cancer associated SNP being an eQTL shows indeed stronger evidence and should be favoured in the PRS SNPs selection. Additionally, integrating those data could provide insights into the biological mechanisms involved in lung cancer susceptibility.

An unexpected result from this work was that the association observed between germline variants and mutational burden was revealed in lung adenocarcinomas only and not in lung squamous cell carcinomas. Previous studies have highlighted so far differences between the two subtypes. At the molecular level, Campbell *et al.* showed that LUSC can be closer to other squamous carcinomas than LUAD [212]. Also, the effects of several lung cancer susceptibility SNPs have been reported as different between the histological subgroups [175, 80]. However, the chromosome 15q25 locus is not one of them and is strongly associated with smoking behaviour, a common risk factor for both subtypes. While the differences observed between

the two cohorts might be specific to the TCGA dataset, another reason could be that even if smoking is a common risk factor, different smoking behaviours are observed between the two cancer types (See Introduction section 1.1). In the TCGA lung cancer cohorts, those differences were observed, with a difference in the proportions of never smokers in LUAD and LUSC. Hence, the selection of samples based on their histology or on their smoking status could induce collider bias. This bias generally occurs when an exposure and an outcome impact a third variable, the collider. However, it can also result from a non-representative selection of samples, *e.g* when the exposure and the outcome variable drive the selection of the samples in the study [213]. This bias could explain the absence of association in the LUSC samples and more broadly could impact such analyses where only cancer cases are considered.

Using multiple MR tests, a causal effect has been identified between cigarettes per day and mutational burden. However, pleiotropy has also been detected by the MR-Egger test. In addition, leave-one-out analysis indicated that the causal effect was driven by one genetic instrument located on the *CHRNA5* region (chromosome 15q25 locus). Further analyses restricted to the chromosome 15q25 locus would be needed since the detected pleiotropy might be related to this locus. This locus has been indeed identified so far as associated to different traits (lung cancer, smoking traits, chronic obstructive pulmonary disease (COPD)). Although the causal genes involved have not been fully identified, two main genes have shown stronger evidence: *CHRNA5* and *Iron Responsive Element Binding Protein 2 (IREB2)* [214]. In the work of Bosse *et al.*, the knock-down of *IREB2* has been associated with an increase of DNA damages in human lung fibroblast [214]. As the biological mechanisms impacted by the chromosome 15q25 locus remain unclear, the use of other omics datasets could bring new insights. As a preliminary analysis, gene set variation analysis (GSVA) was performed on expression data. The expression profiles of the two rs10519203 homozygote groups (See Figure 4.10A-B) were compared to identify mechanisms underlying their molecular differences. The analysis revealed a differential activity for two hallmarks of cancer: "Enabling replicative immortality", in which telomeres play a central role [7], and "Genome instability and mutations" involving DNA repair mechanisms (*q-values* of 0.021 and 0.098 respectively, Figure 4.10C). In line with these preliminary results, the HumanBase online portal [211] reports that *CHRNA5* might interact with DNA repair genes in the lung tissue (see Figure 4.10D). To confirm the involvement of those pathways and identify others potentially involved, other characteristics of the tumors, like the copy number variations, rearrangements, methylation profiles or other variables derived from molecular data like immune context variables [64], could be explored and explain

the observed heterogeneity. For that purpose, a supervised method like PLS-DA or DIABLO could be used in order to identify molecular features discriminating the two homozygote groups for rs10519203 (See Introduction section 1.4).

Currently, few studies are combining and taking advantage of both publicly available GWAS and multi-omics data. Nevertheless, there is a wide range of opportunities for such studies. To begin with, the TCGA data processed here could be further explored by investigating the other cancer types individually, *e.g.* one could study the influence and mediation of Body Mass Index (BMI) on breast and colorectal tumors or of alcohol on oesophagus tumors. Based on the same data, another possibility would be to develop a pan-cancer approach. Alternatively, the ICGC and PCAWG projects [55] (See Introduction section 1.2) represent other valuable resources providing whole-genome sequencing data for around 40 cancer types.
Furthermore, while we focused our study on the total number of mutations and Signature 4 mutations in lung tumors, other mutational signatures identified in these cancers could be investigated. Alexandrov *et al.* suggested that the Signatures 2 and 13 could be indirectly resulting from tobacco smoke, *e.g.* as a result of inflammation or indirect outcome of DNA damage [177].

The study presented in this chapter faces several limitations. Firstly, in contrast to the large datasets usually used to explore germline susceptibility, the sample size used here, reaching less than 1000 samples, was relatively small. Depending on the lists of susceptibility SNPs considered for the PRS or the MR analyses, the association with mutational burden varied, indicated either that the association is more complex than expected or that there is a lack of statistical power. Also, the somatic landscape of a tumor can be complex; the interaction between somatic events could make the interpretation of the results more difficult. For example, some somatic mutations or tumor characteristics like Microsatellite Instability (MSI) status can influence the mutational burden of tumors. Such tumor characteristics could bias the association studied here. For theses different reasons, replicating the results in an independent dataset would be needed.

Additionally, our analyses considered only European samples, since the main ancestry represented currently in public databases is the European ancestry. Ducan *et al.* highlighted that 67% of the GWAS studies conducted between 2008 and 2017 were based on European populations and showed that PRS scores derived from those studies under-performed in non-European populations due to differences in genetic architecture and allele frequencies [215]. The lung cancer susceptibility locus on chromosome 15q25 itself is not found in the Asian population. Hence, most of the work presented here does not generalize to other populations. The sample size

and the diversity of genetic studies are though currently increasing and projects specifically aiming at exploring differences between cancers around the world are emerging. One of these projects, the Mutographs project [216], is adapted to study germline-somatic interactions since it attempts to explore, based on WGS, the mutational signatures in five cancer types from diverse populations.

Finally, another limitation of the analyses relates to measurements errors. On one hand, the mutational signatures were attributed based on the WES data, which could lead to uncertainty, especially if the number of mutations in those tumors is low. Thus transforming the number of mutations attributable to Signature 4 from a continuous variable to a categorical variable (presence or absence of Signature 4) might be more appropriate. Also, the Signature 4 which is associated to smoking can be observed in patients exposed to other chemicals like arsenic, benzene or bisphenol [217]. In this study, the presence of the Signature 4 was observed in some never smokers, which could be explained by passive smoking cases, misclassifications or the implication of one of the other exposures previously mentioned. Each case could impact the association tested between germline susceptibility and the Signature 4 mutation burden. Besides, the work of Alexandrov *et al.* on the new mutational signatures [32], highlighted that previous signatures could reflect overlapping signals and be separated. This overlap of mutational processes could hinder the detection of associations between germline events and mutational signatures and complicate their interpretations. On the other hand, the choice of adapted variables to study the smoking trait is not always straightforward. In 2008, Le Marchand *et al.* suggested that cigarettes per day might not be a good measure of smoking dose [218]. They showed, for two variants on the chromosome 15q25, that carriers tended to smoke more intensively (even for the same amount of CPD) and were thus exposed to higher levels of nicotine per cigarette dose. While this observation points out a limitation in our SNPs selection, especially the instruments for the MR analysis, it might explain the pleiotropic effect associated to this loci. In addition to the difficulty to measure the smoking exposure, the heterogeneity of the trait is high (*e.g* smoking depth will impact smoking heaviness in addition to the common CPD feature). Further work considering more diverse and adapted measurements of smoking exposure would thus probably help to disentangle the several impacts of lung cancer susceptibility on the lung tissue in conjunction with smoking. In this context, Wooton *et al.* proposed recently a measure of lifetime smoking exposure by combining multiple smoking traits measurement like smoking duration, heaviness and cessation, in what they called the lifetime smoking index [219]. Instruments for this index could replace the smoking-related instruments used in this work.

4.2.5 Contribution

I performed most of the data processing and analyses presented in this chapter. In terms of data processing, I performed the imputation of the TCGA samples from all cohorts available and co-supervised a master student to automatize and improve the workflow for future reuse. Regarding the analyses, I gathered the TCGA clinical and somatic molecular data from public resources. I combined and harmonized summary statistics coming from a GWAx on lung cancer (performed by Dr. Atkins) and on smoking-related traits (from the GSCAN study). Based on the previously mentioned datasets, I finally performed the PRS and MR analyses.

Chapter 5

General discussion

The work presented in this thesis took advantage of omics datasets and integrative analyses in order to shed light on distinct lung cancer types. Chapter 2 characterized the LNEN samples at the molecular level and identified relevant molecular subgroups. In particular, the atypical carcinoids, usually described as having an intermediate survival, were stratified in two groups with poor and good prognosis respectively. In Chapter 3, the generation of a molecular map of LNEN samples has been described, and data sharing and reuse were promoted by providing both the transcriptomic data underlying the pan-LNEN molecular map as well as the pipelines required to reproduce the analyses or to process new data following the same workflows. Finally, in Chapter 4, germline and somatic data of NSCLC cancer samples were integrated to investigate the germline somatic interactions in these cancers. While associations between germline susceptibility to lung cancer and mutational burden in lung tumors were identified, they were driven by tobacco smoking susceptibility SNPs and pleiotropic effects were detected, suggesting more complexity. While the main results of this thesis have been discussed in each chapter, the following paragraphs expand on some of the limitations and possible extensions of the different studies as well as on how those analyses fit in the current and future field of cancer genomics.

5.1 Multiple ways of integrating omics data

The three chapters of this thesis take advantage of integrative analyses to explore omics datasets. In cancer genomics, we can think of different types of data integration. Firstly, multi-omics data measured in the same individuals can be combined. In this thesis, multiple layers of omics data, like expression and methylation, were used to identify subgroups of tumors with specific molecular profiles and to characterize them (chapters 2 and 3). One limitation of this approach though is that the

interactions between the different layers are not taken into consideration. Indeed, as described in the introduction section 1.1, cancer biology is based on complex regulatory networks that span all biological layers, including the genomic, transcriptomic, post-transcriptomic, post-translational and epigenetic levels. Approaches that take into account the biological relationship across omics would be valuable to identify and understand the mechanisms involved. While methods like MOFA (used in chapter 2) can identify sources of variations shared across distinct omics layers, it does not capture the underlying mechanisms, *e.g.* which methylation events regulate specific gene expressions. Such analyses could be expanded by focusing on omics interactions. More and more studies, like the one presented in chapter 4, attempt to decipher how molecular alterations impact each other. For example, correlation analyses between omics datasets can be performed. Calabrese *et al.* identified associations between DNA and RNA alterations, *e.g.* associations between mutations and splicing events or fusions and rearrangements [124]. Another example is the Enhancer Linking by Methylation/Expression Relationships (ELMER) method, which was developed to infer regulatory elements by combining expression and methylation data. Such methods have been used in regulatory networks analyses [220] to have a deeper understanding of the processes of cancer development and progression and to identify cancer drivers.

While integrating different omics layers brings valuable insights on cancer development, they could be combined as well with non-omics data like clinical data. Firstly, such integration is helpful to determine the value of the molecular groups that could be identified by omics studies. In chapter 2, for example, we used the histopathological classification to identify clinically relevant subgroups of pulmonary carcinoids, whose survival data were contrasted. Besides, as described in this thesis, aside from the endogenous molecular events, cancer can result from exogenous processes like environmental exposures and lifestyle. In the work presented in chapter 4, we attempted to better understand the influence of lung cancer risk variants on somatic events in conjunction with the smoking exposure. While an association between germline variants, related to lung cancer and smoking behaviours, and mutational burden was identified, pleiotropy was also detected and its origin remains unclear. In this case, a detailed description of the patient's *exposome* would be beneficial to understand the causal pathways of these associations. The description of the *exposome* could go from a more detailed characterization of exposures (*e.g.* smoking intensity, duration, type of smoking, other exposures like air pollution) to further information on clinical characteristics such as the development of other chronic diseases (*e.g.* COPD in the case of lung cancer), and measurements of lung functions

(*e.g.* Forced Expiratory Volume (FEV)). Those information could, in our case, help to disentangle the pleiotropic effects identified. However gathering comprehensive clinical data is challenging [221, 54]. Patients can move around different centers in the care system where medical records of patients are often not organized uniformly across countries and even inside a single country. Thus, along with the molecular data, improving the availability of such data in the next years would be required.

Finally, another way of performing data integration consists in combining datasets from different studies performing the same measurements on different sets of samples. This integration type has the advantage not only to increase the sample size of the study but also to enable to compare samples coming from distinct tumor types in the case of pan-cancer studies. In chapter 2, contrasting the pulmonary carcinoids with the LCNEC samples has enabled to identify the supra-carcinoids having the histopathological characteristics of the pulmonary carcinoids but the molecular features of the LCNEC samples. This observation highlighted aggressive pulmonary carcinoids and supported the hypothesis of a potential link between low and high grade LNENs. Hence, completing the data integration with other cancer types could lead to the discovery of new entities and generate new hypotheses on those potential links. Performing cross-cancer studies can also allow to identify and better understand common carcinogenesis mechanisms [222]. Indeed, across cancer types, key dis-regulated pathways overlap, and the pleiotropic nature of cancer alterations, *i.e.* their ability to influence multiple pathways and diseases, could be utilised to better understand cancer mechanisms. In risk prediction, it has been shown that combining GWAS summary statistics on multiple traits can improve risk prediction tasks in contrast to single-trait analyses [223]. In this context, the work presented in chapter 4 on germline-somatic interactions could be extended. Based on exomes or genomes, a GWAS on complex phenotypes like somatic features (*e.g.* mutational burden or DNA repair molecular signatures) could be performed using a cross-cancer cohort to explore further the influence of germline susceptibility on molecular events. While the sample sizes of somatic studies are so far not comparable to the one used in classical GWAS, future genomics projects will help to reach the sample size needed to answer such questions.

In the work presented in this thesis, the data integration performed relied on computational analyses based on different machine learning methods, both supervised methods with the use of random forest and regression analyses, and unsupervised methods like dimensional reduction techniques. The complexity of the

data to integrate, including the difficulty of gathering complete and accurate clinical data, the high dimensionality and the biological complexity of the data, may require though to expand on those methods.

5.2 Expanding on machine learning methods

People often think of machine learning or more broadly artificial intelligence as methods able to outperform what humans can do. However, these algorithms face biases similar to those impacting humans decision making [224]. For example, as mentioned in this thesis, due to the curse of dimensionality and the limited size of datasets, it is not easy to evaluate and describe rare events. Also, the methods are used on data gathered, and in the case of supervised methods, labelled by humans, they are thus prone to errors. In chapter 2, a random forest classifier has been trained on molecular data to recognize the histopathological classification of LNEN samples. However, it is known that the classification of those samples is imperfect and misclassifications can occur. To avoid training a model on uncertain labelled data, one option could be to use semi-supervised learning. This category of machine learning lies between supervised and unsupervised analysis since both labelled and unlabelled data are considered. The idea is to first train the model on the labelled data, the model is then applied on the unlabelled data and confident predictions are iteratively incorporated in the labelled dataset to retrain a model that should be improved [102, 225]. In the case of the LNEN data, samples for which a consensus was reached among pathologists could be used as labelled data and the samples whose classification is more uncertain as unlabelled samples.

The high dimensionality of genomics data, as well as the complexity of cancer biology, complicate the analysis and interpretation of such data. In this thesis, we used machine learning algorithms that can capture complex structures in the data, like random forest and dimensionality reduction methods. In the past years, deep learning, a branch of machine learning methods, has been commonly used in computer vision and text processing to learn more complex features from the data. In the field of cancer research, deep learning has been applied mostly on images like histopathological slides [106]. Considering the increased number of genomics studies, such methods could though also be applied to genomics data. In the case of smaller studies like the LNEN studies presented in this thesis, the applicability of deep learning algorithms is however limited. One solution though, would be to take advantage of the large databases and apply what is called transfer learning. This method consists in reusing existing models, usually trained on large datasets.

The parameters of the model are not randomly initialized since retrieved from a pre-trained model. This way, the basic features in common between the training dataset and the dataset, on which the model is transferred, do not need to be learned again. In the deep learning field, repositories storing such models have been created and called model zoos. One of them, named Kipoi [106], is dedicated to deep learning models applied on genomics datasets and is in line with the promotion of open access research. Such repositories will probably grow with the large amount of omics data that will be available in the future, enabling the use of deep learning methods on raw omics data like expression or methylation. Note that other solutions exist; examples of transcriptomic data transformation to 2D images have indeed been already proposed to enable the use of deep learning method for cancer classification [226, 227] and could be extended to multi-omics analyses. Also, apart from classification problems, other research questions in genomics could be explored using deep learning approach, *e.g.* GWAS variants prioritization [228].

While they can capture complex patterns in the data, often, machine learning methods, in particular deep learning, are criticized for being difficult to interpret, they are considered as "black box" models. Indeed the biological value of the patterns identified by the methods can be explored only if the model is interpretable. Multiple strategies already exist for that purpose [229]. Some methods perturb the model inputs and explore the impact of the perturbations to identify important features. Other methods, such as random forest, for example, can directly provide feedback on the most discriminating features by inspecting the model's parameters. Our work in chapter 2 could be improved in this context. Indeed, using random forest, molecular subgroups of atypical carcinoids with different prognoses were identified. However, considering the limited number of samples and to limit overfitting, a leave-one-out analysis was performed. This led to the generation of one classification model for each sample and hence complicated the interpretation of the final classifier. Gathering more samples, especially the atypical carcinoids, would allow us to replicate the analyses and would be more adapted for the extraction of important features characterizing the two groups of atypical carcinoids.

Integrative approaches, as well as computational methods like machine learning, have been broadly applied in the past years to study lung cancer genomics on both the germline and somatic levels. However, due to the complexity of cancer and genomics data, which has been revealed to be more complex than initially thought, multiple challenges remain.

5.3 Future challenges in lung cancer genomics studies

The three chapters of the manuscript focused on lung cancers, which are tumors that have been relatively well characterized over the past decades. From a germline point of view, lung cancer GWAS have identified susceptibility loci associated with lung cancer overall but also with lung cancer subtypes and with samples from different smoking status. In terms of risk prediction, lung cancer has been described however as one of the cancer types for which PRS have a limited added value for risk assessment in comparison to existing criteria [230], which is probably due to the modest heritability of lung cancer [158] as well as the strong association of the disease with smoking. The sample size needed to explain 80% of the GWAS heritability for this cancer type has been estimated to 1,000,000 cases [158]. Integrating existing and new large datasets would increase the statistical power and lead to new susceptibility loci identification. However, such scale for lung cancer cases analyses will be difficult to reach in the near future. In the meantime, it has been suggested that the next steps to undertake in lung cancer genetics would be to identify the causal genes and pathways involved in the disease and to improve our understanding of the underlying complex biological networks. Integrative and causal approaches like the ones presented in chapter 4, contrasting germline susceptibility and somatic molecular events, can help in this direction.

On the somatic level, omics studies of lung cancers have revealed clinically relevant molecular profiles of lung cancers. However, most of these studies, including the ones presented in this thesis focused on bulk tumor cells analyses. The molecular profiles observed in those data result from a mixture of heterogeneous cells, that can come from the tumor but also its microenvironment and thus might not be representative of all cells issued from the clonal evolution of the tumor. Bulk tumor cells studies are thus limited to explore tumor evolution and the influence of tumor microenvironment on cancer development. Recently, the number of spatial studies exploring multiple pieces of the tumor has increased to reach those goals. In the case of lung cancers, the Tracking Cancer Evolution through Therapy (TRACERx) project, launched in 2014, has been developed. The project is a longitudinal study following more than 800 lung cancer patients and aiming at exploring NSCLC evolution. The analysis of the first 100 cases have already provided valuable insights on lung cancer progression and foreshadow a better comprehension of NSCLC evolution in the next years [231]. Also, single-cell analyses have been developed to explore omics data at the cell level and study the heterogeneity among millions of cells. This large data scale is thus particularly appropriate to apply data integration and methods like machine learning, in comparison with studies on bulk tumor cells

whose low sample sizes lead to multiple challenges.

In addition, while common lung cancer tumors have been so far relatively well characterized, rarer forms of lung cancers would benefit from further investigations. Chapter 2 performed a molecular characterization of the rare lung neuroendocrine tumors based on multi-omics data identifying distinct molecular clusters of pulmonary carcinoids. Even if this study corresponds to the largest multi-omics datasets on LNEN samples, the sample size is still limited, and the results require validation, in particular the discovery of the supra-carcinoids, composed of only six samples. Further characterization of those tumors would require to gather more pulmonary carcinoids samples. In the near future, the LNEN molecular groups identified will be further characterized in the context of the rare cancers genomics initiative to confirm their existence as well as to explore their potential link with other lung cancer types. In line with the latter objective, the molecular map described in chapter 3 is currently being updated with the integration of other lung tumors, LUAD and LUSC samples, and will be further completed with the data from the new LNEN study.

Finally, the etiology of several lung cancers still needs to be explored. Indeed, while smoking is the strongest lung cancer risk factor, lung cancers are also detected in never smokers. For that purpose, a better description of the patient's *exposome* is required. As described at the beginning of this chapter, lung cancer genomics studies would benefit from the integration of omics with non-omics data such as clinical information. One step in that direction would be to take advantage of the patient's electronic health records that would in future studies be more frequently available. Although the analyses of such data raise challenges, like data heterogeneity and complex data types (*e.g.* unstructured data), the use of methods, like machine learning, could help to integrate them in genomics studies.

As mentioned in the previous paragraphs, multiple lung cancer genomics studies would benefit from larger samples sizes. In the following sections, we describe how future genomics projects around the world could assist and complete existing studies and how the analyses presented in this thesis could be translated in the context of those new data.

5.4 The establishment of larger omics datasets

The extension of large research projects

Across the three chapters of this thesis, one limitation has been the access to a limited number of samples. In the two first chapters, the fact that LNEN cancers are rare is an obstacle to genomics data collection. The final sample size might not provide a complete picture of the molecular profiles existing in these cancer types. In the third chapter, the germline somatic interactions in LUAD and LUSC cancers were explored. While those lung cancer types are common and the use of the TCGA data allowed to reach almost 1000 samples, the genetic effects of common variants tend to be weak. As such, larger samples sizes will be required to examine weak genetic effects in this context extensively. As described in the introduction section 1.2, after the TCGA initiative, larger genomics projects like the ICGC or the UKbiobank have emerged and are currently being extended for research purposes. The ICGC has launched a new project called the Accelerate Research in Genomic Oncology (ICGC-ARGO) project aiming at applying sequencing techniques on more than 100,000 cancer samples. The project has already gathered more than 50,000 donors and plans to provide curated and complete clinical data to accompany the genomics data, which, as mentioned previously, is usually a challenge for large genomics projects [54, 232]. Also, the 500,000 UKbiobank samples are going to be whole-genome sequenced to explore further germline genetics. Finally, the coordination of several national health programs mentioned in the next paragraph could provide additional access to genomics data. Those expansions give us the opportunity to reproduce the studies presented in the manuscript on larger datasets in order to replicate the results and to complement them with further analyses.

The development of clinical projects

In the past years, the genomics studies in academic research have influenced the way cancer patients are managed and have paved the way to national genomics programs designed for the use of such data routinely in the clinic in order to advance the field of precision medicine [233]. Indeed, multiple national genomics projects have started worldwide. In France, for example, in the context of the "France médecine génomique 2025" project, two sequencing platforms were launched in 2017 and the sequencing of around 40,000 genomes is expected every year [234]. In the UK, the 100,000 genomes Project launched by Genomics England in 2012 planed to sequence 100,000 patients, including patients with cancer or rare diseases [235]. This initial

goal was reached in 2018 and the plans were expanded to the sequencing of one million genomes in the following years. Both projects aim at integrating genomics data analyses in health care by taking advantage of what has already been learnt in the era of genomics as well as to provide new insights on human diseases like cancer. In the context of the latter goal, the Genomics England Clinical Interpretation Partnership (GeCIP), which forms groups of researchers, has been established to study specific domains to improve our understanding of the different diseases and the application of genomic medicine in clinical care [236].

Translating omics research results to the clinic

The generation of several genomics studies provides valuable knowledge on oncogenesis processes that can be leveraged to inform individual or groups of newly sequenced samples. One example of such an application is the Cancer Genome Interpreter (CGI). The CGI is an open platform that gathers and takes advantage of genomics information from thousands of already sequenced tumors to explore mutational patterns and identify clinically relevant driver alterations [237]. Similarly, the approaches and results presented in this thesis could be used to inform the clinical care of newly sequenced samples.

In the first chapter, taking advantage of expression and methylation datasets, molecular subgroups of pulmonary carcinoids have been identified. Those clusters, characterized by distinct prognoses, presented added value to the current histopathological classification. Future analyses validating and further characterizing these clusters could allow to identify potential biomarkers and inform future classification of LNEN tumors. This analysis provides an example of how genomics analyses can guide the diagnosis of new patients. However, in the context of the current and future clinical projects mentioned previously, efforts concentrate on one omic layer, the genome. Analyses like the one presented in chapter 2 are based though on transcriptomes and methylomes and suggest that incorporating the use of those data in clinical settings, such as the national programs described previously, would be beneficial for the patients.

In chapter 3, we suggested, in the discussion, that molecular maps could be used as a reference to project new samples. In a future where thousands of samples will be sequenced in the clinic, molecular maps could be explored to contrast the projected samples with the tumors already characterized. This could be particularly valuable for cases that are difficult to classify and even lead to the identification of misclassified samples. Also, using the molecular and prognostic characteristics

of the reference tumors, to which the projected sample matches, could provide insights on how the tumor might evolve and respond to different treatments. While the main objective of the future clinical national programs is to use the patient's genome to identify known driver alterations to guide the clinical care, the previous examples show how exploiting the similarities and differences between previously characterized samples could guide the clinical care of new patients.

In chapter 4, PRS were used to provide a measure of lung cancer risk for each individual. While in this study, PRS measures were correlated with mutational burden to explore germline-somatic interactions, it could also be applied in a clinical context for cancer prevention and early detection. Indeed, PRS can identify patients with high risk for which early and regular screening protocols could be beneficial. While we discussed that the use of PRS in lung cancer has, for now, limited value in addition to smoking related criteria, PRS could be developed to stratify specific categories of samples, *e.g.* among each smoking categories. In addition, PRS are still often criticized for their lack of interpretability, which harms the credibility of these methods for application in the clinic [195]. Approaches similar to the one proposed in chapter 4 could be helpful to uncover the mechanisms underlying germline susceptibility and deal with interpretability issues.

Challenges of translational research

One difficulty raised by the application of research findings in the clinic is the selection bias. Indeed, researchers need to keep in mind that the samples recruited in an academic setting may not be representative of the whole population or of the patients that are going to be treated in the clinic, due to various sources of bias. Firstly, bias can occur due to inclusion criteria under which samples are selected in cancer genomics studies. For example, samples under treatment or with previous disease conditions can be discarded. However, in a clinical setting, patients with antecedents might be frequent. Hence, the results found in genomics research could sometimes not replicate in a clinical setting. Surgical resections can not be performed on all patients with cancer, and the use of biopsies is often not adapted to genomics studies since they have poor cells content. Those samples are thus often not represented in genomics studies. Also, samples included in research projects are often collected at diagnosis and more clinically advanced cancers might be underrepresented. Finally, clinical genomics cohorts might themselves introduce biases since the studies might focus on specific patients (*e.g.* patients that do not respond to conventional treatments, patients whose diagnosis is difficult or rare cancer cases). Such cohorts would not be ideal, for example, to generate a reference molecular map

that would be used to characterize newly sequenced data, since the map would not be representative of all tumors. These various biases bring additional sources of heterogeneity in genomics studies and will thus complicate analyses like data integration.

Another major challenge of the implementation of cancer genomics datasets is to increase diversity in the data to diminish disparities in cancer research. This challenge is often mentioned when considering germline studies involving, for example, GWAS or PRS analyses. Indeed, most of the current GWAS studies were performed on individuals from European ancestry. The identified associations and derived measures, like PRS, do not generalize to other ancestries. The application of these tools in the clinic is thus limited since it would not be beneficial for a large proportion of the population. In addition, diversity is also a challenge to consider in studies exploring the somatic landscape of tumors. In omics studies attempting to perform molecular profiling of tumors, like the ones presented in chapter 2 and 3, the sample's ethnicity is not always considered. Nevertheless, it has been shown that the molecular profiles of tumors vary between ancestries and that those differences can bias the results. Carrot-Zhang *et al.* showed, for example, that samples from African origins harbour fewer mutations in the *VHL* and *PBRM1* genes in renal cancer and that ancestry can be an important confounder in genomics studies [238]. A large proportion of omics studies, including the ones presented in this thesis, focused so far on European samples. Considering that molecular features are being proposed as biomarkers for clinical use, the inequalities observed in the data composition might translate into disparities in clinical care. Including samples from diverse ethnicities in the genomics projects will not though be enough to overcome the current disparities. Genome references data used in a large number of omics studies were also built mostly on European participants and would need to be updated [239]. Finally, the lack of diversity in the datasets impacts also the interpretation and the reuse of computational models like machine learning models. Indeed if the models are trained on a biased and non-representative training set, it will hardly generalize to the whole population [224].

The enlargement of academic genomics research projects and the development of national clinical genomics projects provide multiple opportunities for data integration and foreshadow the ability to work on datasets with millions of samples in the future. For this purpose, data sharing and open access research across the scientific community worldwide is though required. The next section covers some of the solutions and guidelines developed so far to achieve this goal.

5.5 Sharing resources for genomics data analyses

In chapter 3, which integrated six transcriptomic datasets, we highlighted the importance of harmonized processing workflows for data integration. In order to facilitate the reproducibility of our analyses and data reuse and integration, we provided: i) the pipelines built based on the Nextflow language and Docker containers used for our analyses, ii) the homogenized dataset, and iii) interactive tools like a computational notebook and a TumorMap. The work from this chapter is in line with a broader effort towards open access data and research. Indeed, in the future where datasets with millions of sequenced samples might be generated, tools that can perform reproducible and automatized data processing and analyses will be required.

Recently, multiple initiatives have collected such tools for promoting reproducibility. The nf-core framework has, for example, been created to provide a set of curated and documented pipelines coded using Nextflow [240]. In addition to allowing reproducibility, such pipelines provide the advantage of scaling to most computational environments like cloud servers. These environments enable researchers to analyze large genomics data remotely without having to download and thus duplicate them. Also, as the size of data increases, cloud infrastructures help to reduce execution time and data storage by providing adjustable computational resources for data analyses [150]. The Michigan and Topmed servers [191, 192] that perform imputation on genotyping data in the cloud, as well as the Cancer Genome Collaboratory, a cloud resource developed to analyze the PCAWG data, can be considered as examples [241].

Another way of promoting data re-investigation is to enable data exploration on interactive platforms. To complement the integrated pan-LNEN dataset presented in chapter 4, two interactive tools were used: a Nextjournal interactive computational notebook and the TumorMap portal. The latter enables future users to interrogate the pan-LNEN molecular map directly, *e.g* through basic statistical testing, hence favoring the generation of new hypotheses. In the past years, an increasing number of such interactive tools emerged and encouraged data exploration by a broader research community, since they can be exploited with little to no expertise in computational biology. In the context of large genomic initiatives, a recent example of an interactive platform, enabling online data query, is the Cancer Virtual Cohort Discovery Analysis Platform (CVCDAP) [242]. The platform proposes to perform pre-defined analyses on the TCGA data uniformly processed. With such tools, any researcher can run genomics analyses without even accessing the data.

While multiple tools exist to analyze genomics data, data sharing itself raises

challenges, especially if the data are coming from international collaborations. Restrictions principally due to variable legislation regarding health data sharing and privacy exist. Indeed, sharing genomics data can present several risks. Those data are sensitive since they could be used to re-identify patients included in the research studies and could provide information that the patient might not want to share or that might be used in discriminative ways. To assure the respectful use of genomics data and overcome divergence in legislation, researchers have recently proposed to elaborate an international code of conduct for data sharing [243]. In this context, different initiatives like the Beyond 1 Million Genomes (B1MG) project and the Global Alliance for Genomics and Health (GA4GH) are currently underway to propose common protocols, in Europe and across the world respectively. These projects aim at enabling data sharing across countries as well as developing an analysis framework for the forthcoming large genomics data while following legal and ethical guidelines [244, 245]. The Collaboratory cloud resource mentioned previously, for example, has been developed in compliance with the GA4GH guidelines and has enabled researchers around the globe to access and analyze thousands of genomes. In Europe, 21 countries signed a declaration aiming at sharing 1 million human genomes by 2022 [233]. This data sharing will operate through a European law on data privacy, the General Data Protection Regulation (GDPR), which took effect in 2018 and was an attempt to harmonize personal data protection regulations in the European Union. In the context of genomics data, it imposes, for example, pseudonymization, *i.e.* that samples included in research studies should not be re-identifiable without any additional information. Together the regulations and guidelines mentioned in this paragraph will allow researchers to share appropriately, integrate and analyze omics data with samples sizes that could not have been reached so far and thus promise important discoveries in cancer research in the next years.

To conclude, the work presented in the manuscript has been developed around integrative and computational analyses applied to different omics datasets in order to: i) unveil the molecular diversity of the lung neuroendocrine tumors, ii) take advantage of independent LNEN transcriptomic datasets to increase sample size as well as contrast tumor molecular profiles with the generation of a molecular map, and iii) explore the interplay between germline susceptibility to lung cancer and tumor mutational burden. Potential applications of the results described in this thesis

in a clinical setting were discussed and range from risk prediction to tumor classification and prognosis inference. Finally, while the results and methods used encounter several limitations related to small samples sizes and data complexity, each analysis could be enhanced in the context of larger genomics studies and clinical programs whose data are going to be shared across the international scientific community in the next years.

Appendix A

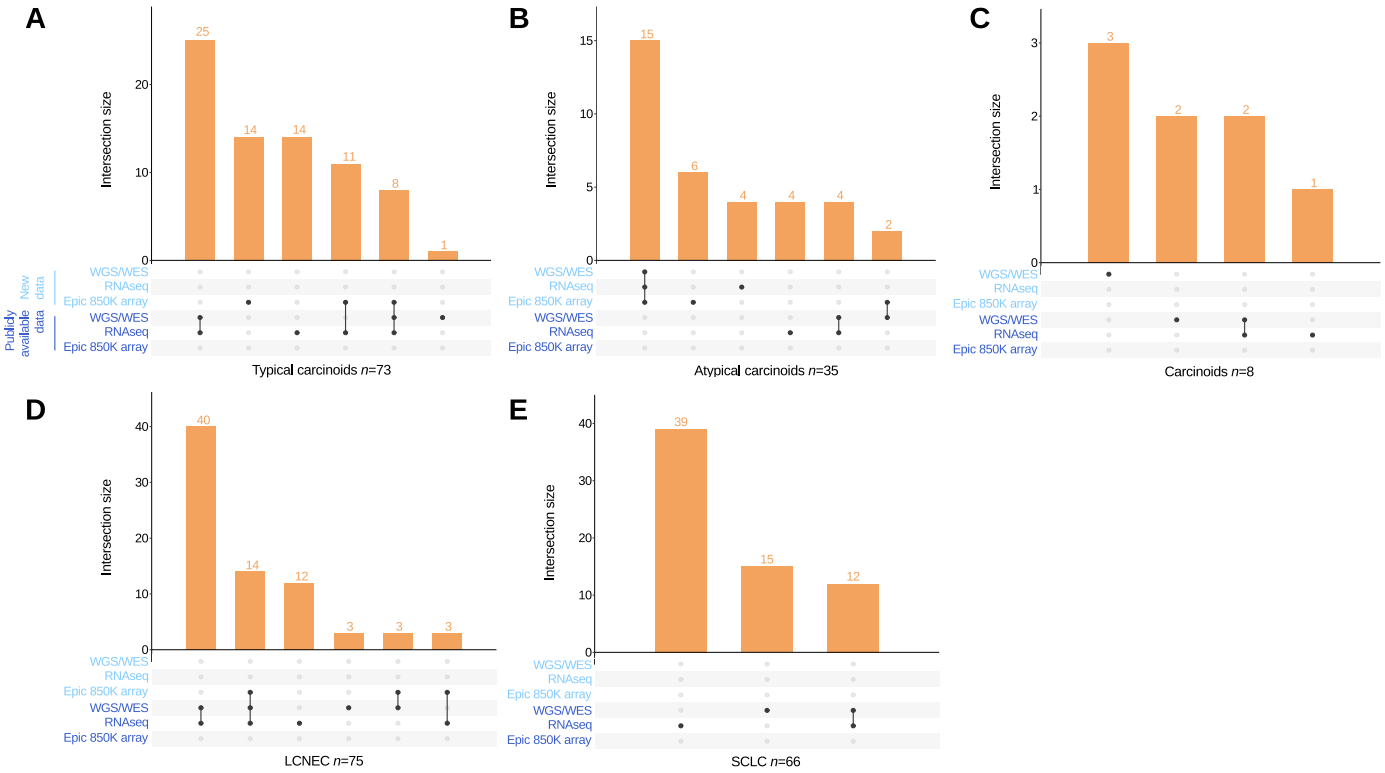
Appendix A

A.1 Supplementary material from Article 1: Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids

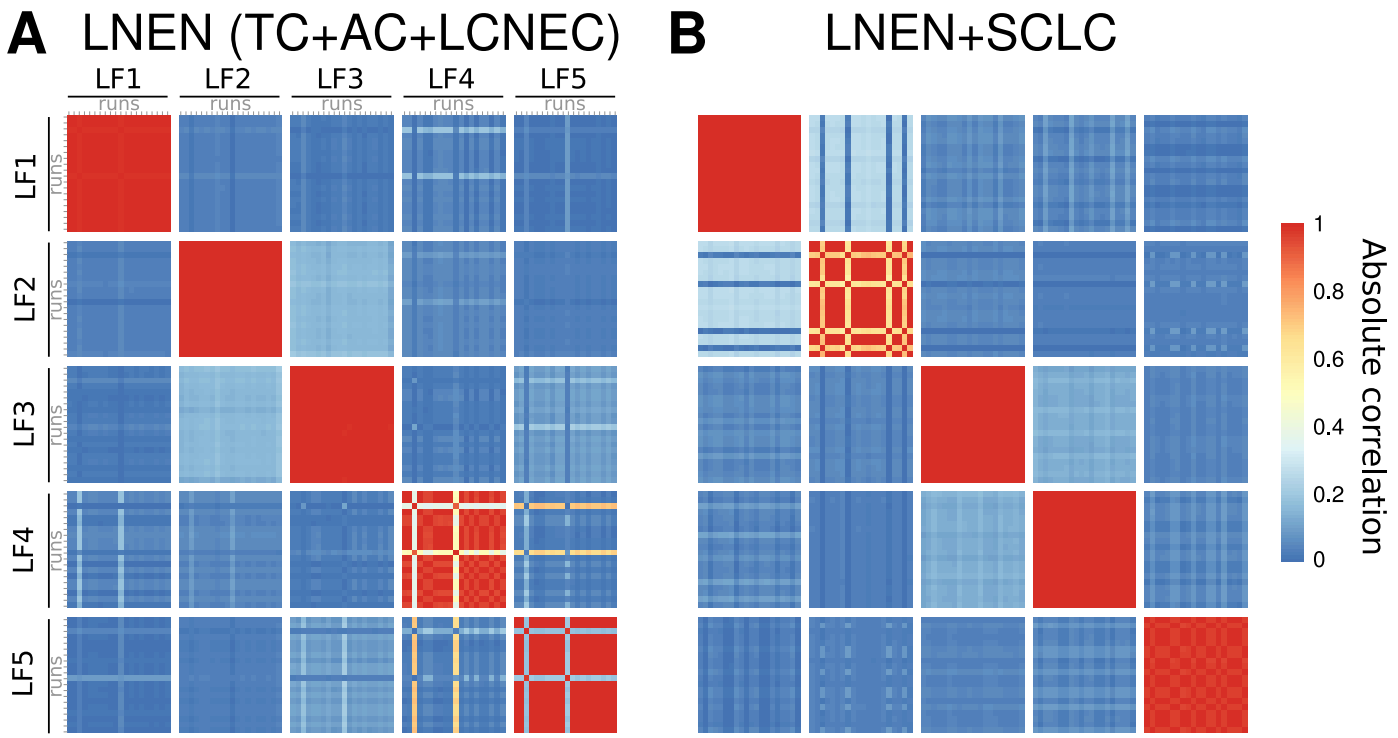
Supplementary Information

Integrative and comparative genomic analyses identify clinically relevant groups of pulmonary carcinoids and unveil the supra-carcinoids

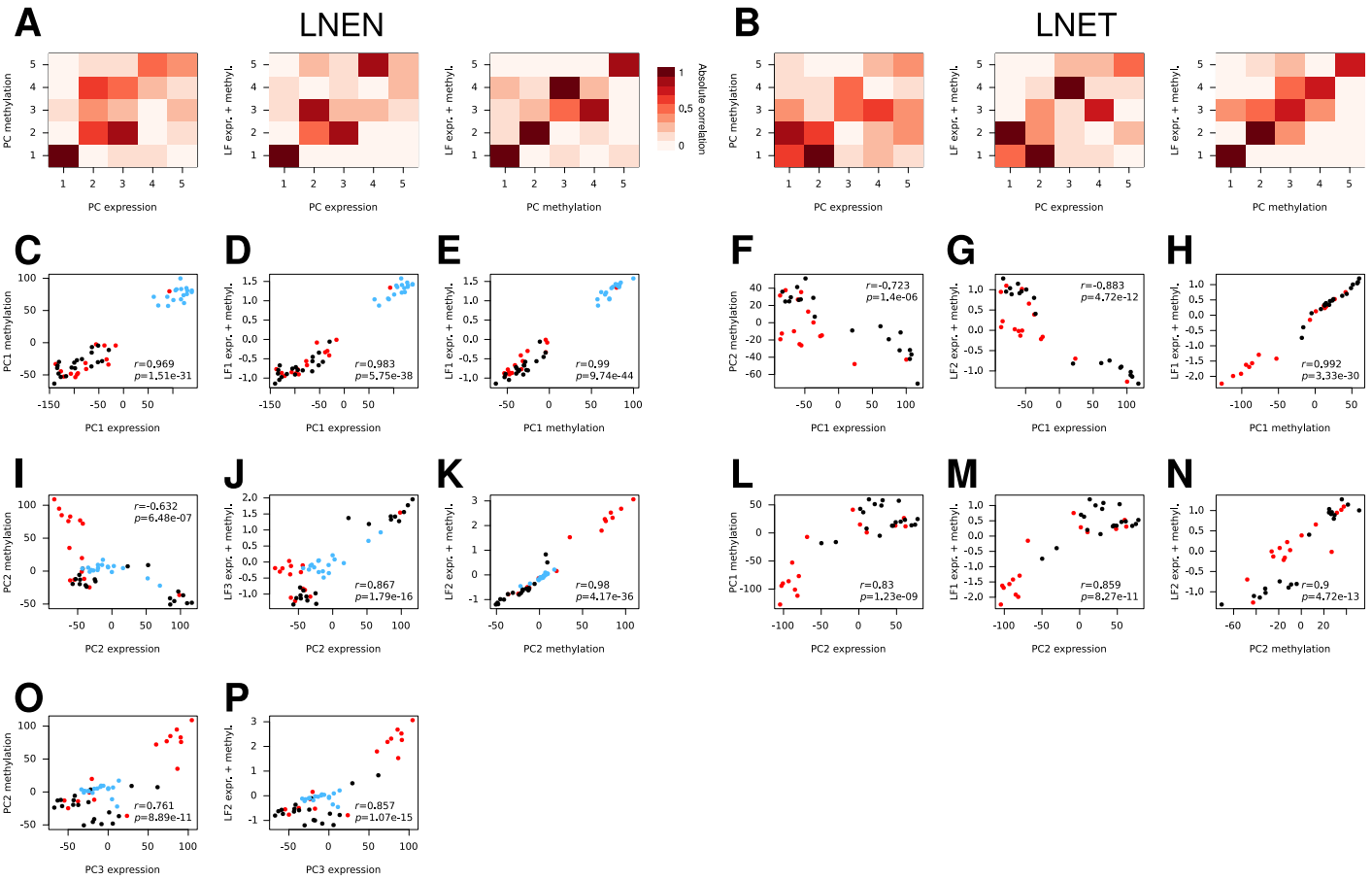
Alcala *et al.*



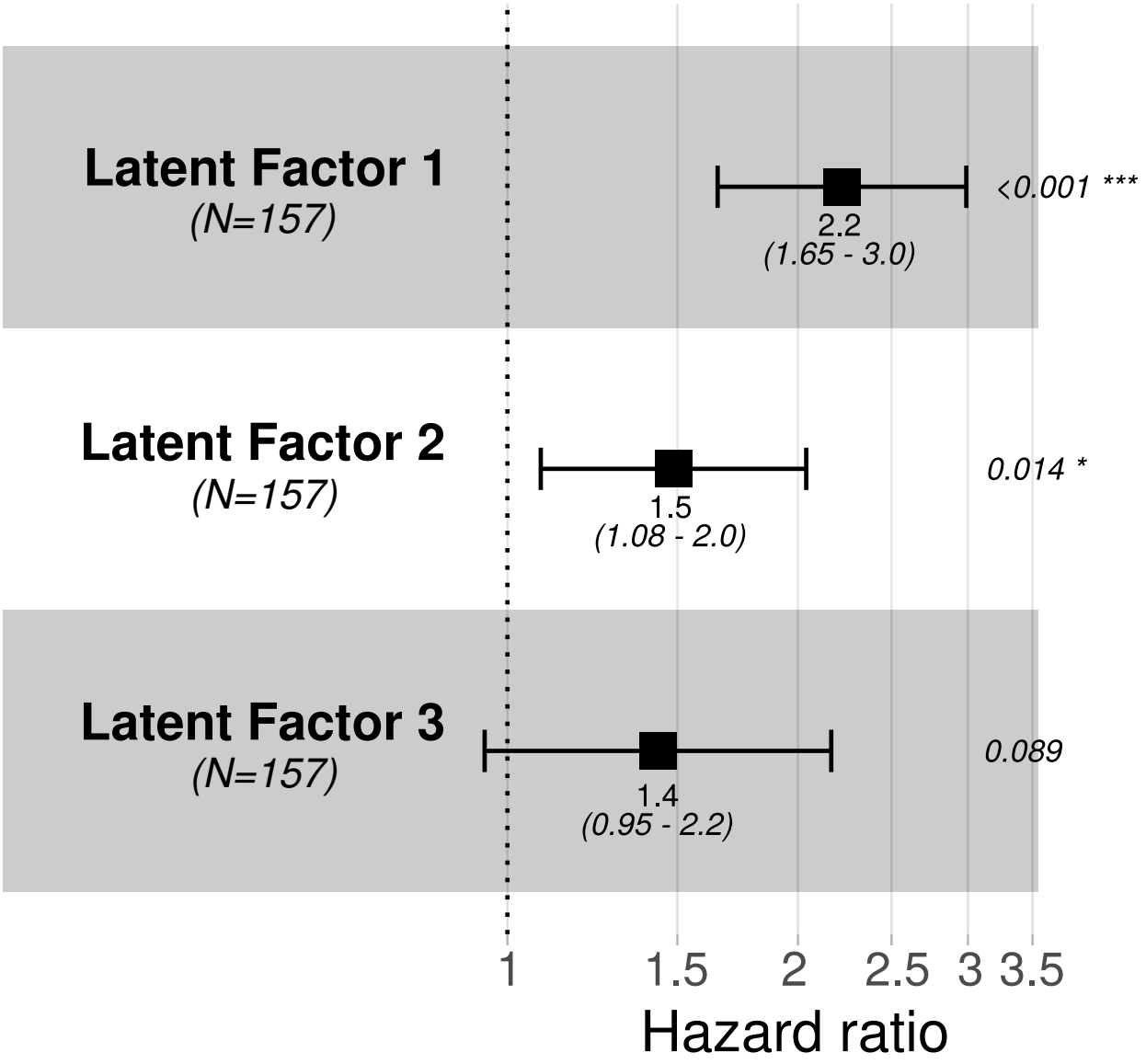
Supplementary Figure 1 Overview of the multi-omic experimental design for LNEN samples. Overview of the number of samples with whole-genome sequencing (WGS) or whole-exome sequencing (WES), RNA-sequencing (RNA-seq), and Epic 850K methylation arrays (EPIC 850K array), for (A) typical carcinoids, (B) atypical carcinoids, (C) carcinoids, (D) large cell neuroendocrine carcinoma (LCNEC), and (E) small cell lung cancer (SCLC). In all panels, new (light blue) and publicly available (dark blue) data are mentioned separately. The total number of samples (*n*) are indicated next to each cancer type. Data necessary to reproduce the figure are provided in Supplementary Data 1.



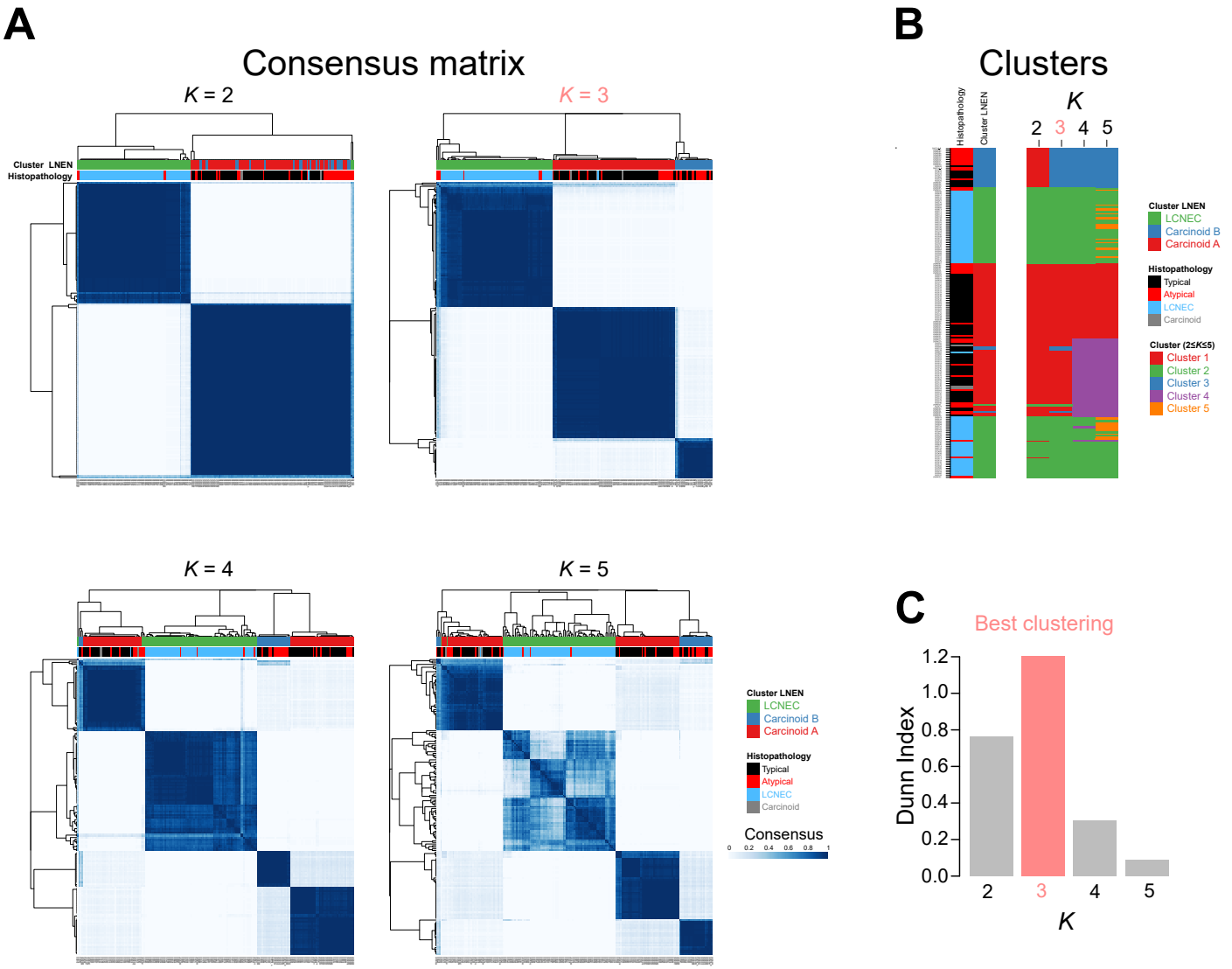
Supplementary Figure 2 Robustness of the MOFA latent factors presented in Figure 1A. Each panel corresponds to the matrix of Pearson correlation coefficients between latent factors (LFs) from 20 replicate MOFA runs. Rows/columns correspond to a single LF from a single MOFA run; rows/columns are clustered by LF (from 1 to 5), and ordered by run number (from 1 to 20) within a cluster (100 row/column in total). Colours represent the strength of the absolute correlation (red for high correlation, blue for low correlation). A) Correlation between LF across runs for MOFA run on all LNEN samples (the best run among the 20 is presented Figure 1A and Supplementary Figure 13B). B) Correlation between LF across runs for MOFA run on all LNEN and SCLC samples (the best run among the 20 is presented Supplementary Figure 13A). In all panels, the red colour on the diagonal and the blue colours off-diagonal indicate a very good robustness of the LF. Data necessary to reproduce the figure are provided in Supplementary Data 1.



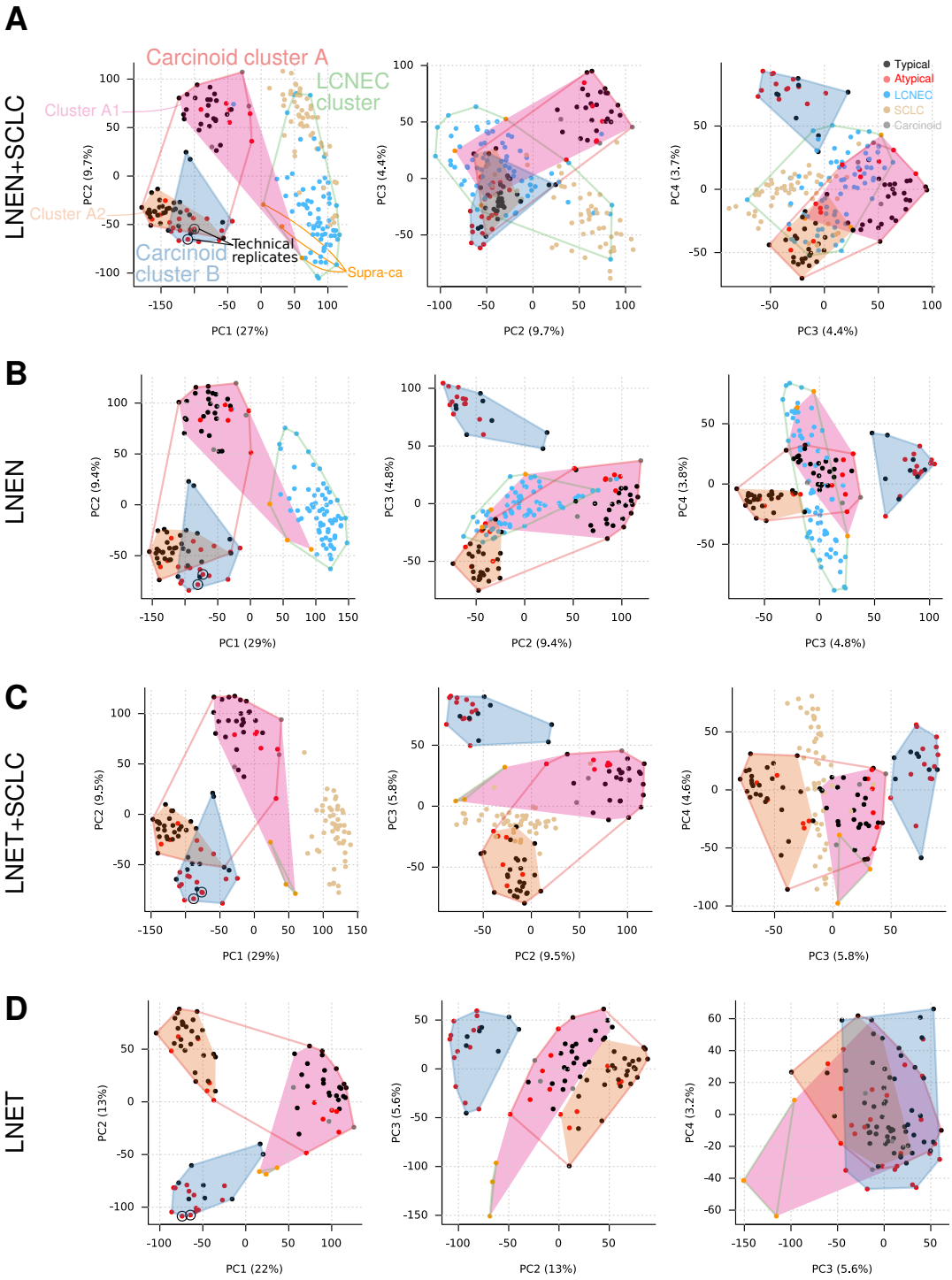
Supplementary Figure 3 Correlations between MOFA latent factors (Figures 1A and 4A) and the principal components of the PCA of expression (Supplementary Figure 6) and methylation (Supplementary Figure 7). Panels (A) and (B) present the correlation matrices between expression and methylation PCA (left), between expression PCA and MOFA (middle), and between methylation PCA and MOFA (right), for MOFA on LNEN samples and LNET samples, respectively. Panels (C)-(P) highlight the strongest correlations from panels (A) and (B) in the form of scatter plots, and display Pearson correlation coefficients and the *p*-values of the associated tests. Atypical, Typical and LCNEC samples are represented in red, black and blue respectively. Data necessary to reproduce the figure are provided in Supplementary Data 1, 2 and 3.



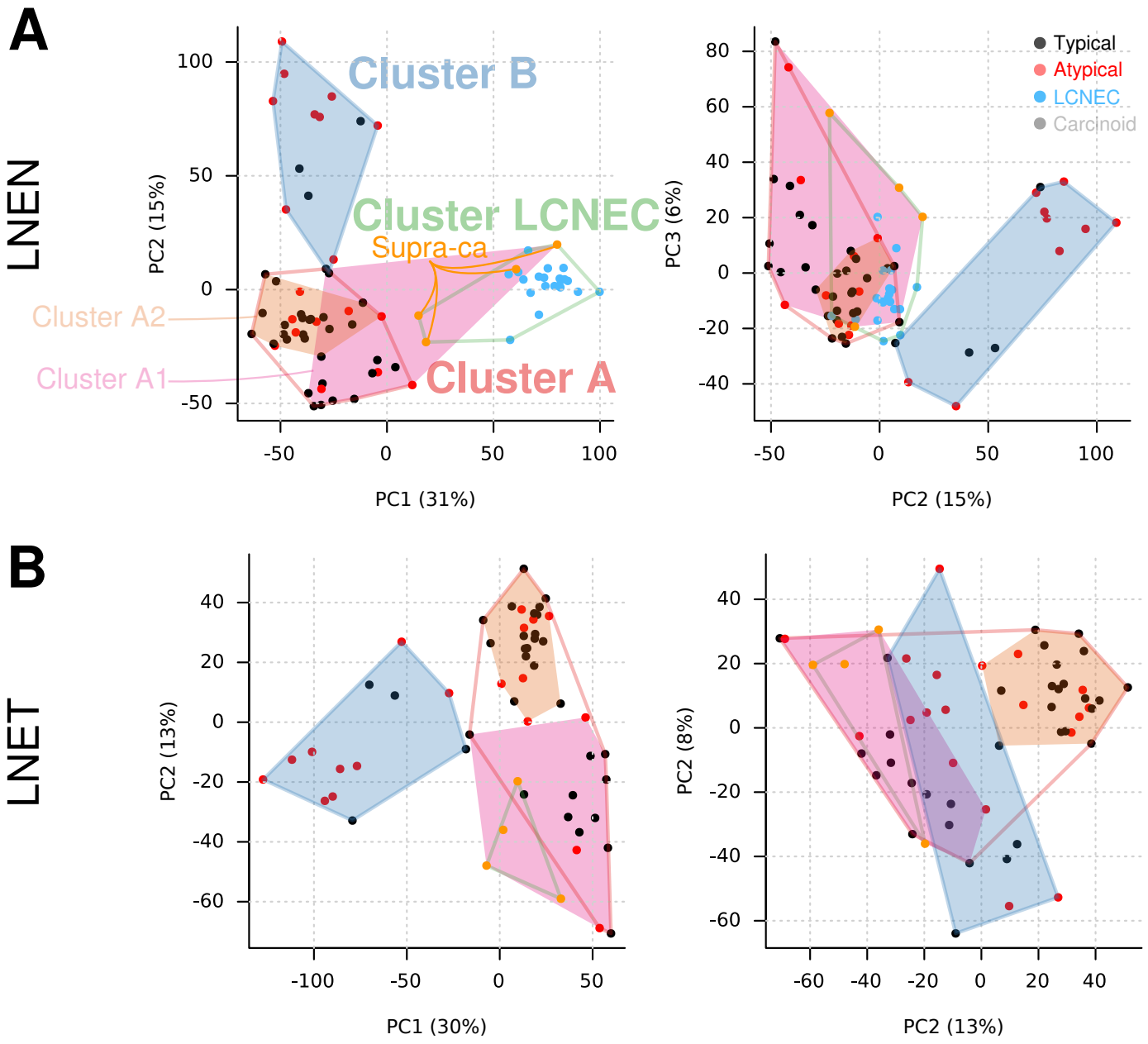
Supplementary Figure 4 Forest plot of the survival analysis based on the first three MOFA latent factors (LFs) of LNEN samples from Figure 1A. Results correspond to a Cox proportional hazards model with coordinates of samples on the first 3 MOFA LFs as continuous explanatory variables. The black box represents estimated hazard ratios and whiskers represent the associated 95% confidence intervals. Wald test p-values are shown on the right; $0.01 \leq p < 0.05$, $0.001 \leq p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. Number of samples (N) for each group is given in brackets. Data necessary to reproduce the figure are provided in Supplementary Data 1.



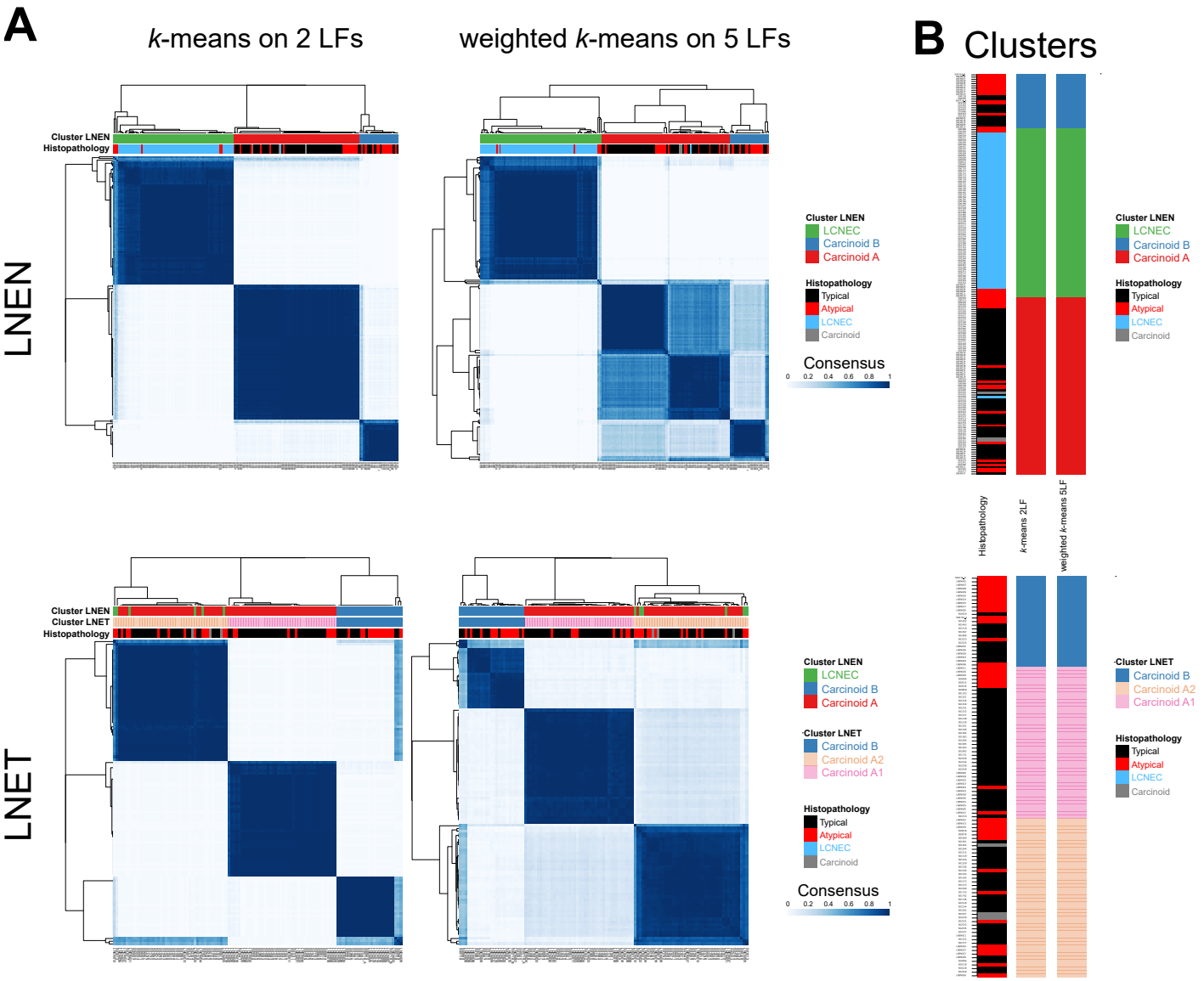
Supplementary Figure 5 Robustness of the consensus clustering of LNENs presented in Figure 1A. A) Heatmap of the consensus matrix for four numbers of clusters *K*; cluster memberships and histopathological types are reported above the columns, and the dendrogram represents a hierarchical clustering. B) Cluster membership as a function of *K*. C) Clustering quality metric (Dunn Index) for each value of *K*; the best clustering according to the metric is highlighted in pink. Data necessary to reproduce the figure are provided in Supplementary Data 1.



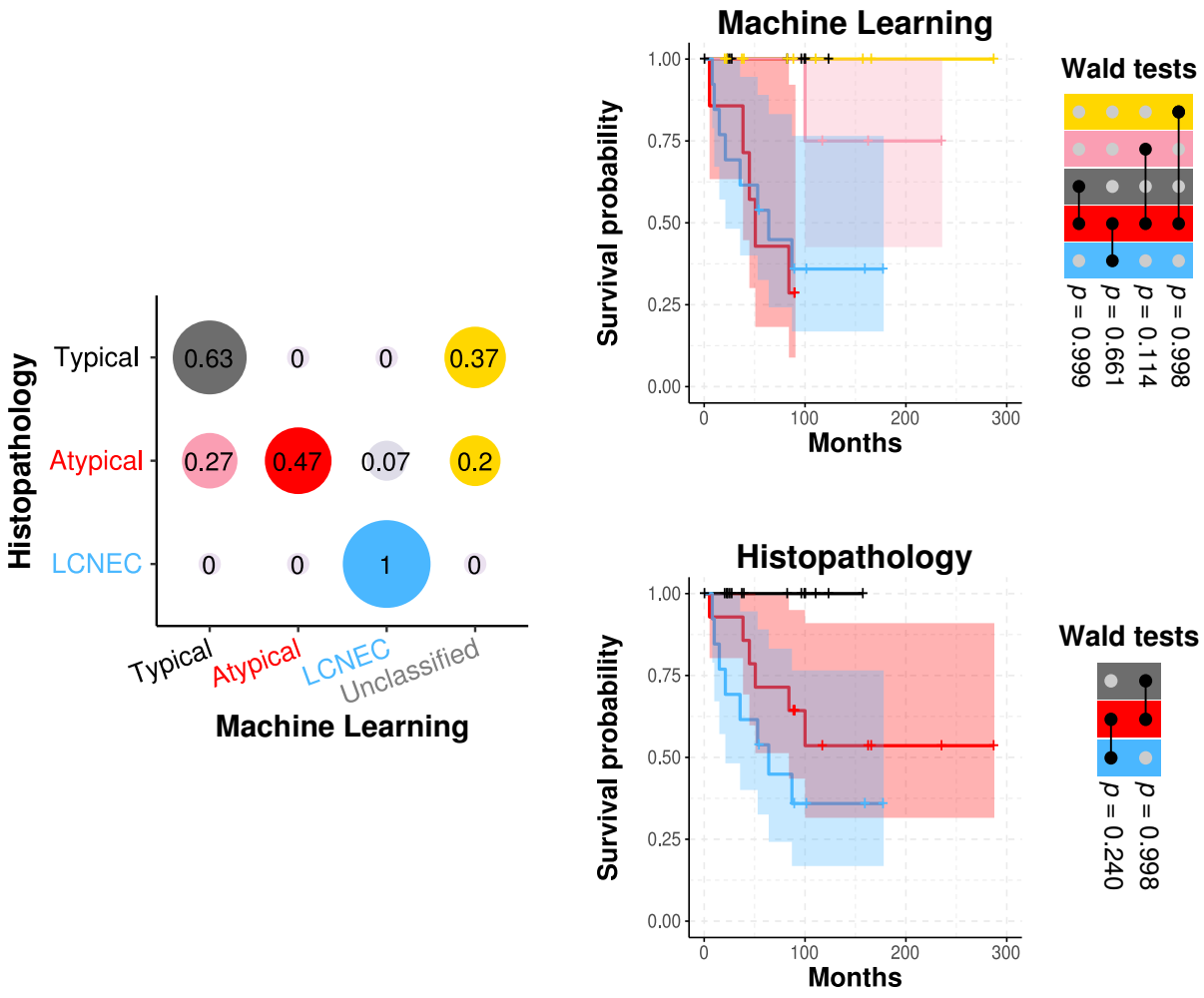
Supplementary Figure 6 Principal Component Analysis (PCA) of transcriptome data. A) PCA of transcriptomes of typical and atypical carcinoids, LCNEC (i.e., LNEN), and SCLC. B) PCA of transcriptomes of typical, atypical carcinoids, and LCNEC (i.e., LNEN). C) PCA of transcriptomes of typical, atypical carcinoids (i.e., LNET), and SCLC. D) PCA of transcriptomes of typical and atypical carcinoids (i.e., LNET). On each panel, point colors correspond to histopathological types (black for typical, red for atypical, grey for carcinoids, blue for LCNEC, beige for SCLC) and supra-carcinoids (orange), polygons correspond to the LNEN clusters from Figure 1A, and filled surfaces correspond to LNET clusters from Figure 4A; their shapes correspond to the convex hull of samples from the focal cluster. The two technical replicates are circled in black. Data necessary to reproduce the figure are provided in Supplementary Data 2.



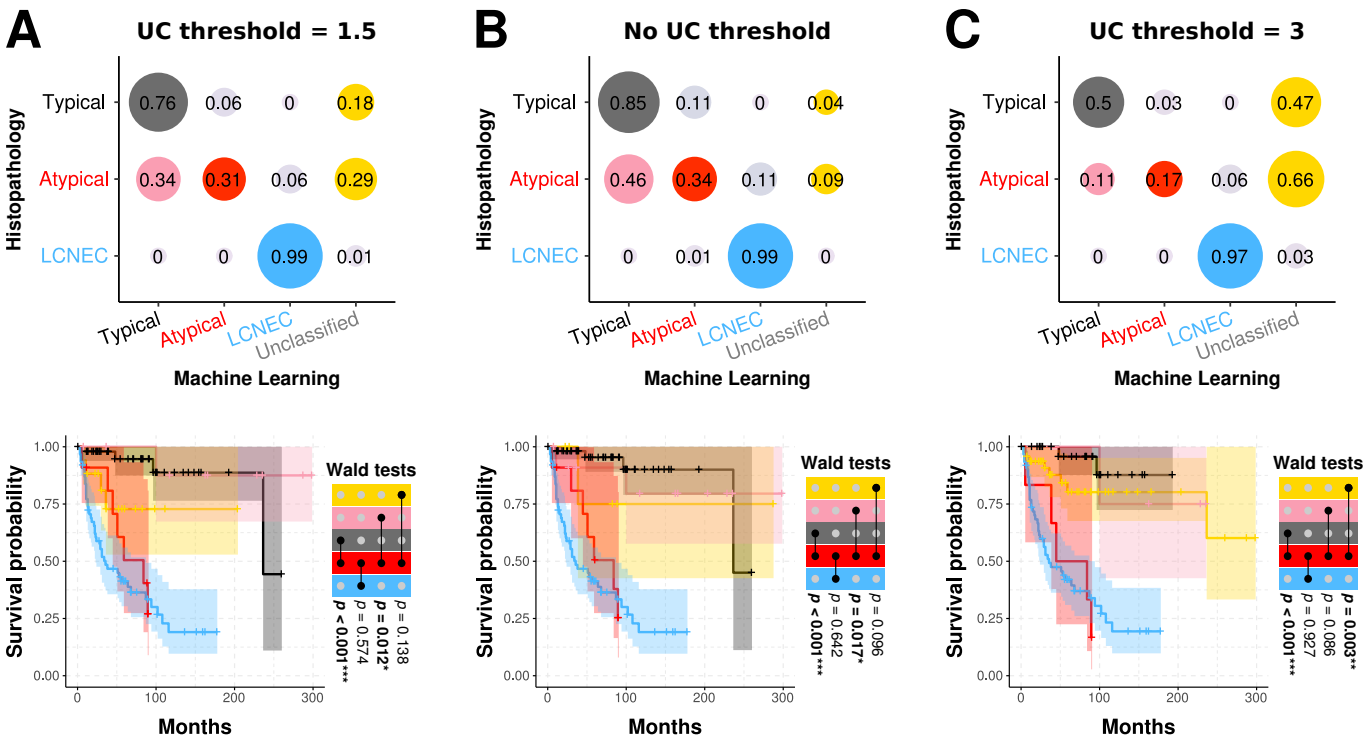
Supplementary Figure 7 Principal Component Analysis (PCA) of the methylation data. A) Analysis of all samples (LNEN). B) Analysis restricted to LNET samples. Figure design follows that of Supplementary Figure 6. Data necessary to reproduce the figure are provided in Supplementary Data 3.



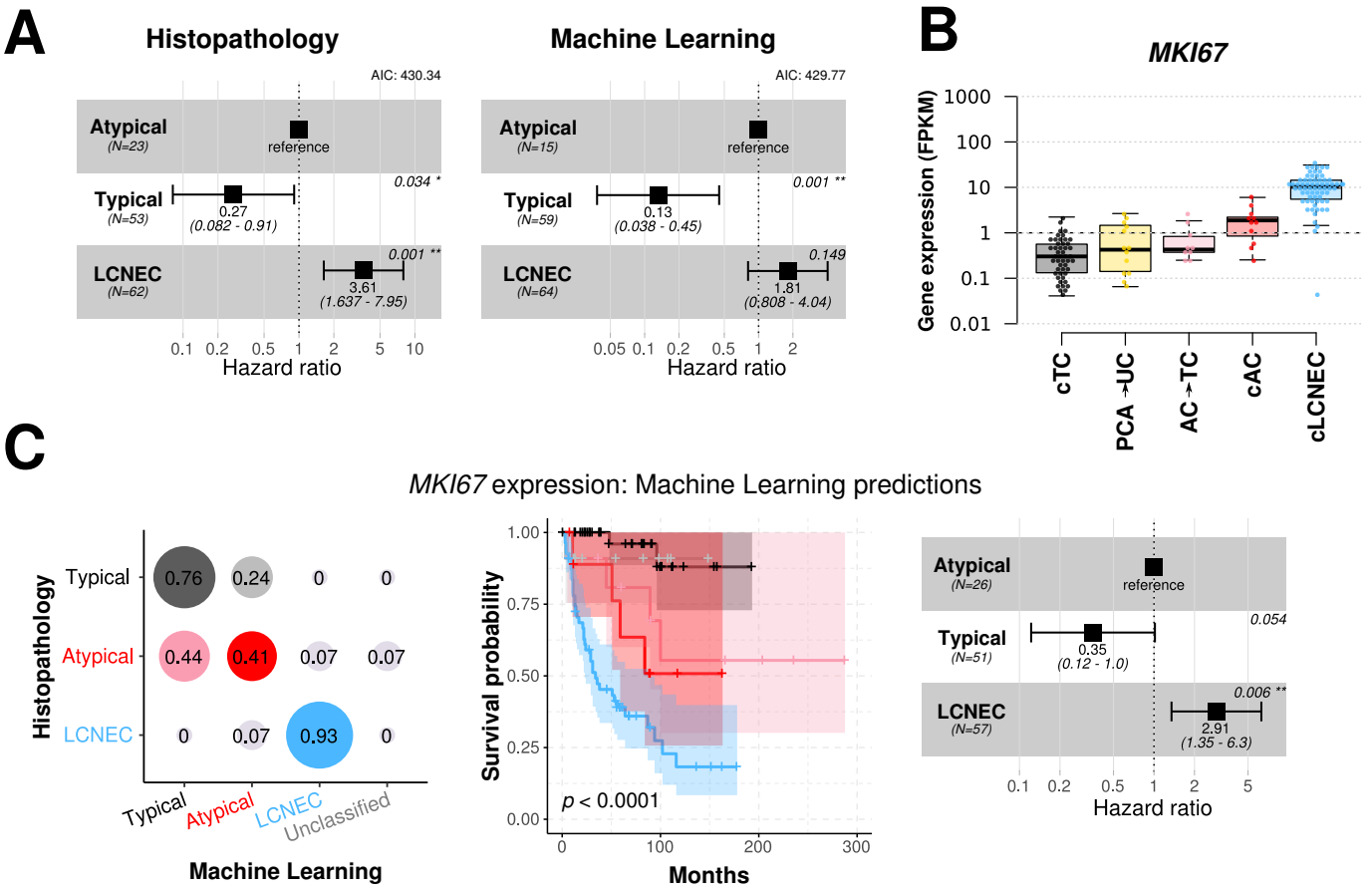
Supplementary Figure 8 Comparison between consensus clustering on MOFA latent factors based on different clustering algorithms. A) First column: copied from Supplementary Figures 5A and 18A; *k*-means clustering using the first 2 latent factors, for LNEN (top) and LNET (bottom) samples. Second column: weighted *k*-means clustering using the 5 latent factors identified by MOFA, weighted by their proportion of variance explained. B) Histopathological type (first column) and cluster membership of each sample, for consensus clustering using the *k*-means algorithm on the first 2 latent factors (second column), and using the weighted *k*-means algorithm on all 5 latent factors (third column). Data necessary to reproduce the figure are provided in Supplementary Data 1.



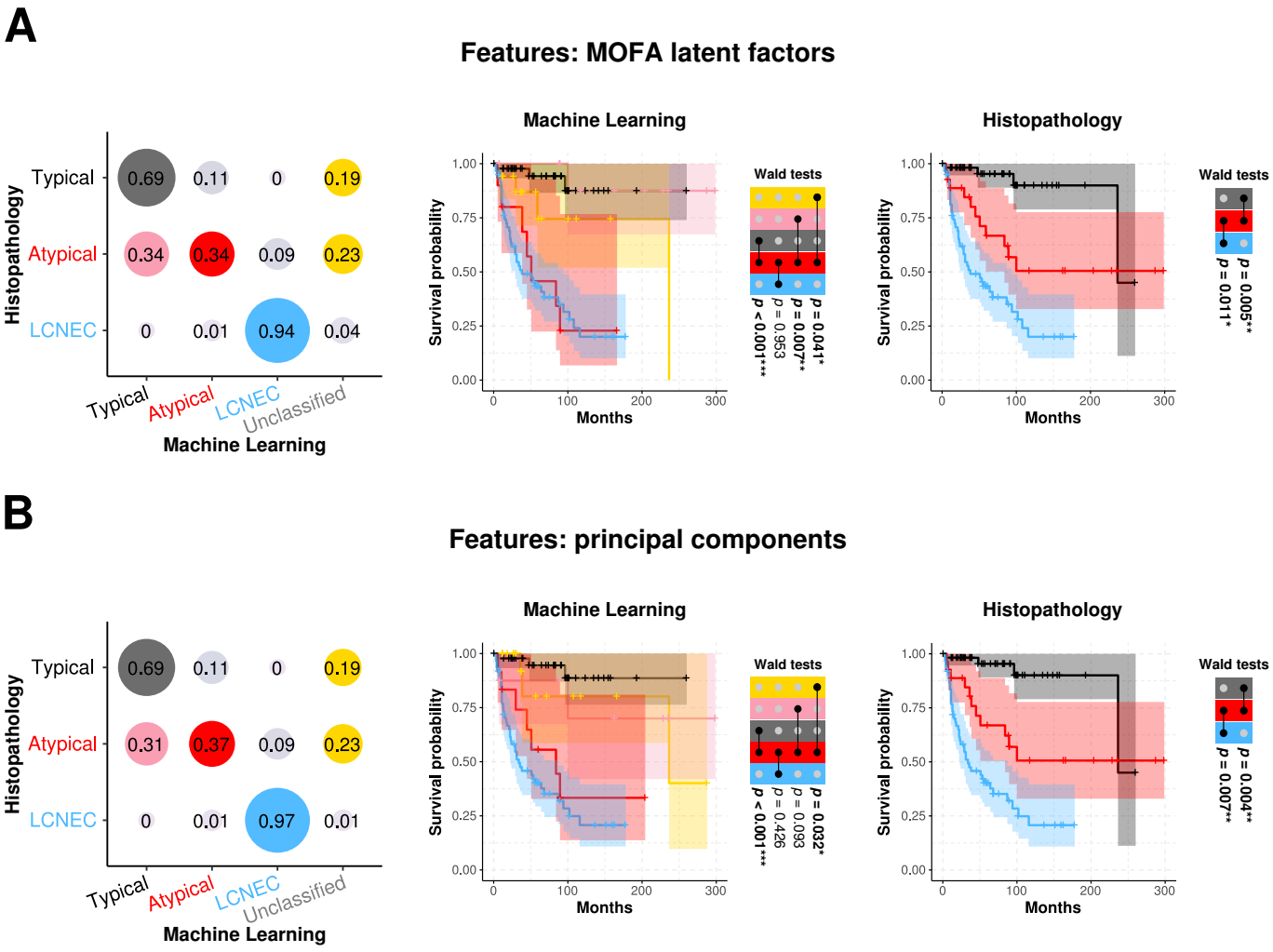
Supplementary Figure 9 Analysis of the ML predictions based on a model integrating expression and methylation data simultaneously. The analysis is similar to that used to produce Figure 1B-C, except that expression and methylation data are integrated simultaneously in the model rather than independently (see Online Methods). Figure design follows that of Figure 1B-C. Data necessary to reproduce the figure are provided in Supplementary Data 1.



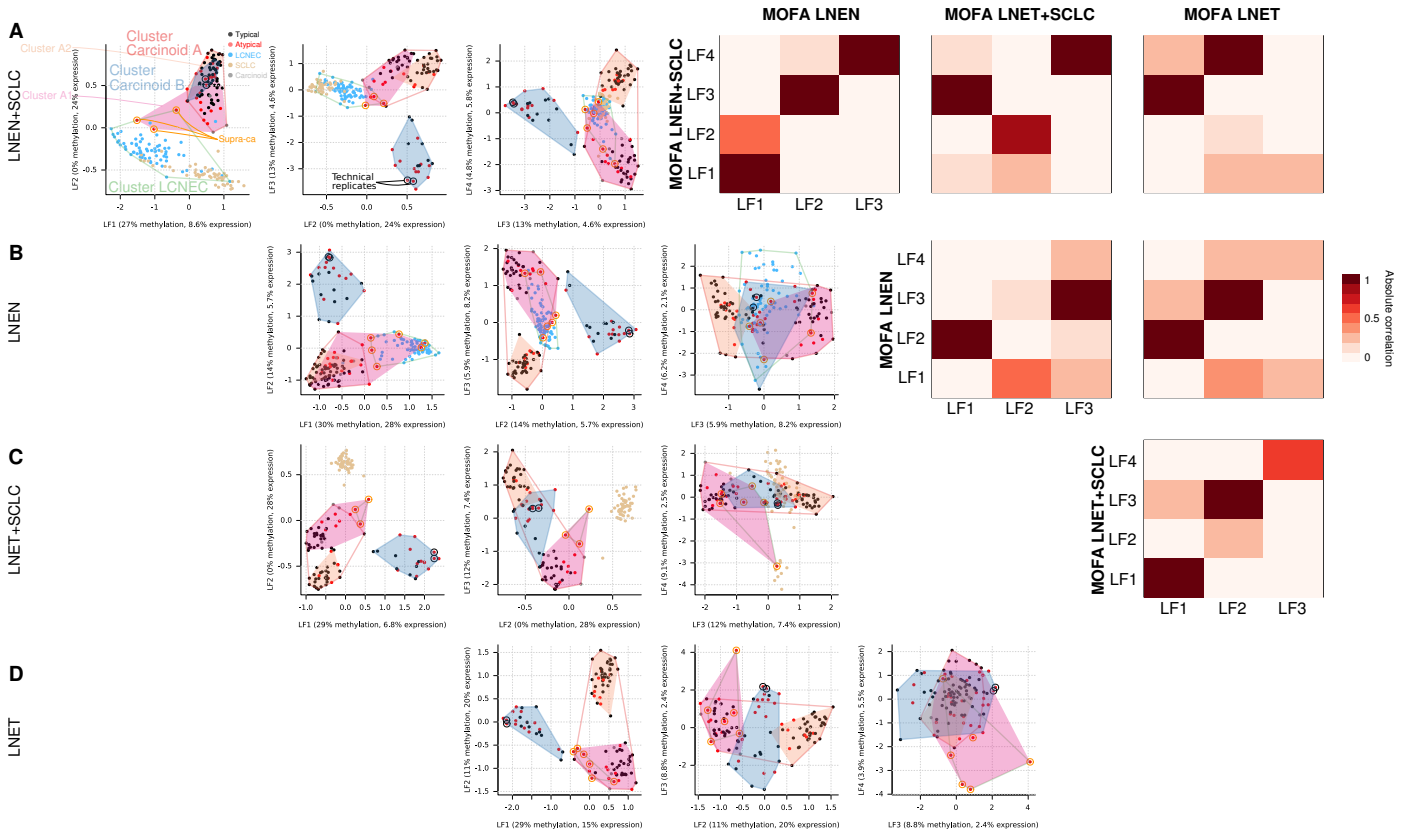
Supplementary Figure 10 Comparison of the ML predictions when applying different thresholds to define the "Unclassified" category. A) Copied from Figure 1B-C for reference. Upper panel : Confusion matrix associated with the ML predictions combined using expression and methylation-based predictions (see Online methods) and a threshold of 1.5 for the definition of the "Unclassified" category. Lower panel: Kaplan-Meier curves of the overall survival of the different ML-predictions groups. B) Upper panel: Confusion matrix associated with the ML predictions combined using expression and methylation-based predictions and no threshold for the definition of the "Unclassified" category. In this case, the only samples predicted as "Unclassified" are the ones with discordant expression-based and methylation-based predictions. Lower panel: Kaplan-Meier curves of the overall survival of the different ML-predictions groups. C) Upper panel: Confusion matrix associated with the ML predictions combined using expression and methylation-based predictions and a threshold of 3 for the definition of the "Unclassified" category. Lower panel: Kaplan-Meier curves of the overall survival of the different ML-predictions groups. For each Kaplan-Meier plot, the colour associated to each group matches that of the confusion matrix in the upper panel. Next to each Kaplan-Meier plot, matrix layouts represent pairwise Wald tests between the reference group (in red) and the other groups, and the associated p-values; $0.01 \leq p < 0.05$, $0.001 \leq p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. Data necessary to reproduce the figure are provided in Supplementary Data 1.



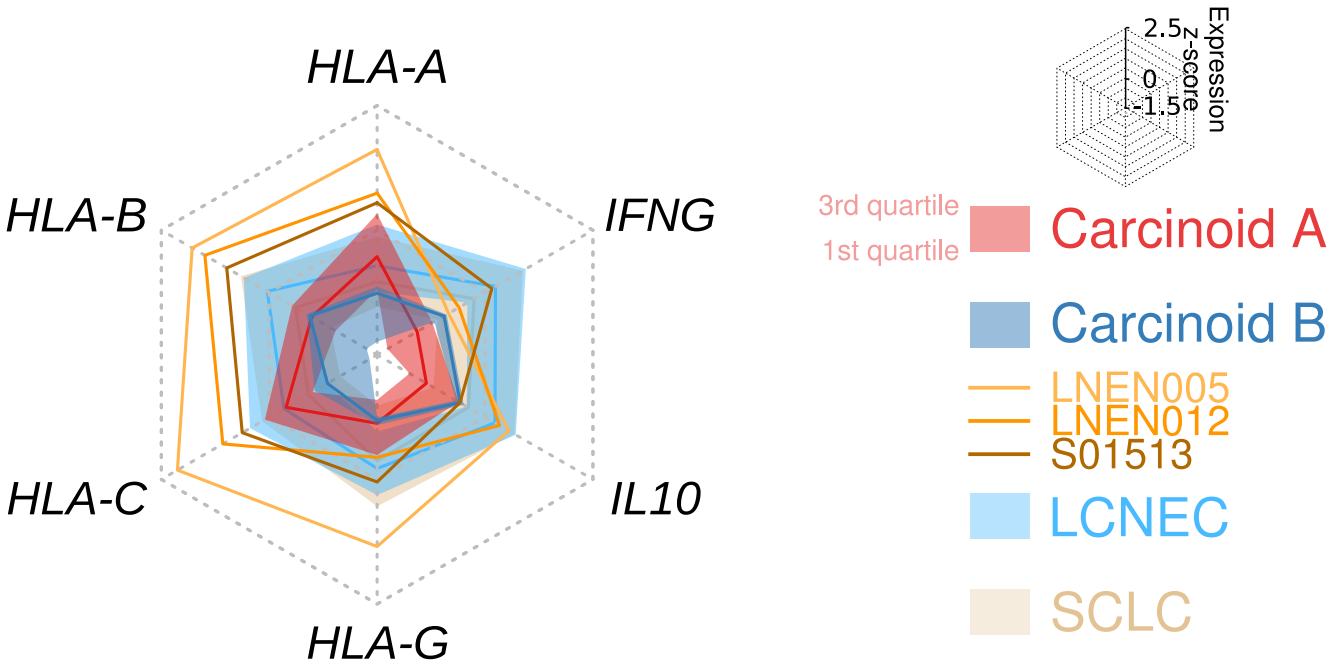
Supplementary Figure 11 Comparison of overall survival based on different classifications. A) Forest plot of hazard ratios of overall survival for two alternative models. Left panel: a model based on the histopathological report. Right panel: a model based on the machine learning predictions from expression and methylation data. For the two models, the same set of 138 samples was considered (see Online methods). B) Boxplot of the expression level (in Fragments Per Kilobase Million; FPKM) of *MKI67* for each prediction group highlighted in Figure 1B. cTC (consensus typical) are typical samples predicted as typical, PCA->UC carcinoids predicted as unclassified, AC->TC atypical samples predicted as typical, cAC (consensus atypical) atypical samples predicted as atypical and cLCNEC (consensus LCNEC) LCNEC samples predicted as LCNEC. Centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. C) Analysis of the ML predictions based on *MKI67* expression only. Left panel: Confusion matrix associated with the machine learning predictions based on *MKI67* expression. Middle panel: Kaplan-Meier curves of the overall survival of the different ML-predictions groups. The colour associated to each group matches that of the confusion matrix (left panel). Right panel: Forest plot of hazard ratios of overall survival for a model based on the ML predictions based on *MKI67* expression. For all forest plots, the black box represents estimated hazard ratios and whiskers represent the associated 95% confidence intervals. Wald test p-values are shown on the right; $0.01 \leq p < 0.05$, $0.001 \leq p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. Number of samples (N) for each group is given in brackets. Data necessary to reproduce the figure are provided in Supplementary Data 1.



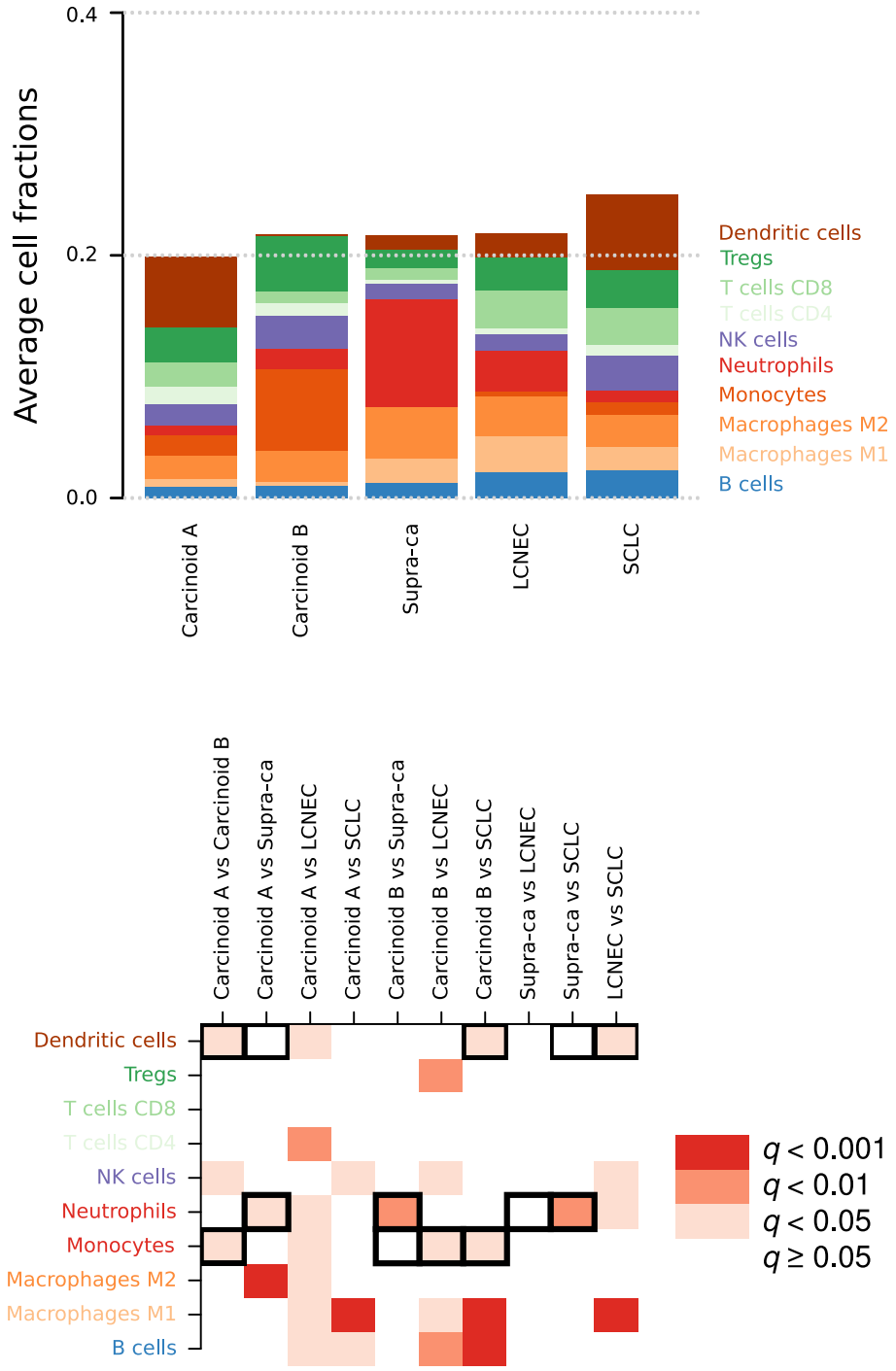
Supplementary Figure 12 Analysis of the ML predictions when considering (A) MOFA latent factors and (B) PCA principal components as features in the classification model. The analyses are similar to that used to produce Figure 1B-C, except that MOFA latent factors or PCA principal components are used instead of expression and methylation (see Online Methods). The MOFA latent factors and principal components explaining more than 2 % of the variance were used in the analysis. The design of each panel follows that of Figure 1B-C. Data necessary to reproduce the figure are provided in Supplementary Data 1.



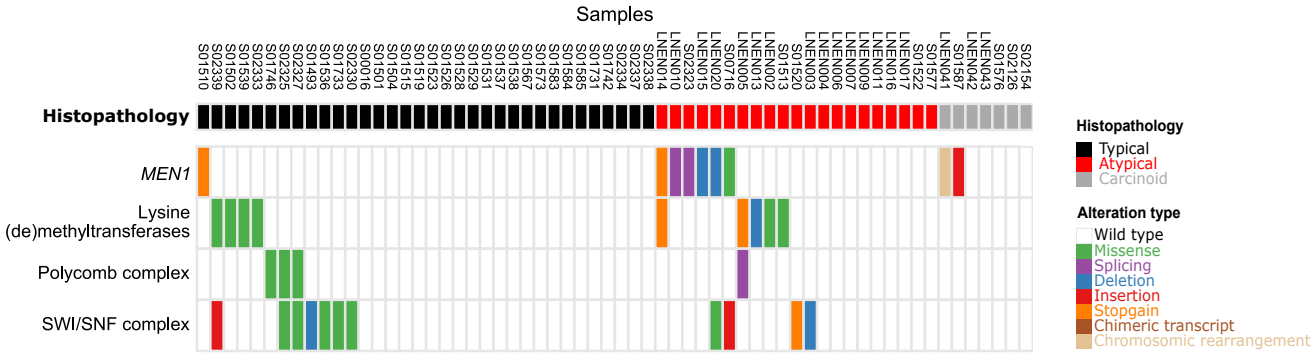
Supplementary Figure 13 Consistency of MOFA across analyses including different histopathological types. A) MOFA of transcriptomes and methylomes of LNEN and SCLC samples. B) MOFA of transcriptomes and methylomes of LNEN samples. C) MOFA of transcriptomes and methylomes of LNET and SCLC samples. D) MOFA of transcriptomes and methylomes of LNET samples. All MOFA plot designs follow that of Supplementary Figures 6 and 7. Images on the right correspond to Pearson correlation coefficients between latent factors (LF) from different analyses (e.g., MOFA LF1 from panel A with MOFA LF2 from panel B, where colors correspond to the strength of the absolute correlation; light: weak, dark brown: strong). Data necessary to reproduce the figure are provided in Supplementary Data 1.



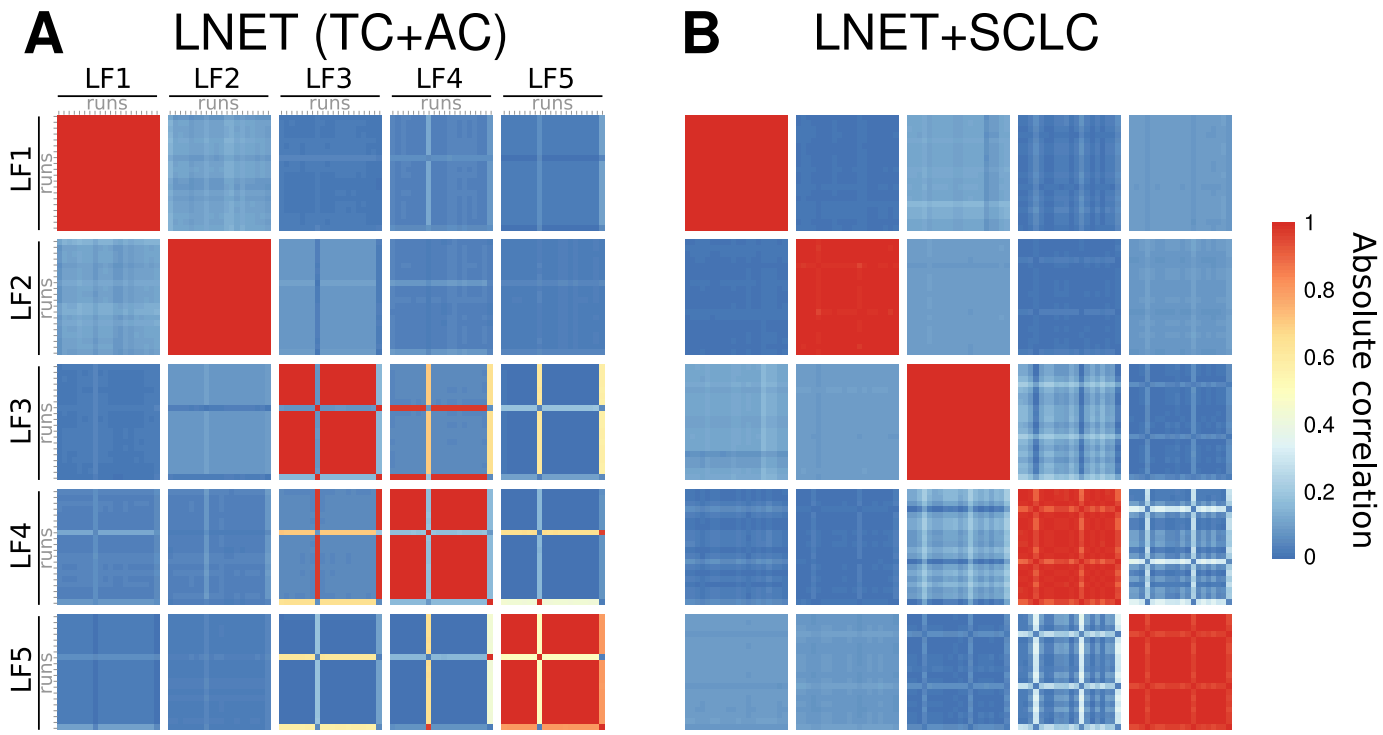
Supplementary Figure 14 Radar chart of the expression levels of HLA class I and related immunostimulatory genes as a function of their molecular group. Expression levels are expressed in *z*-score; the different groups correspond to the LNEN molecular clusters (Carcinoid A, Carcinoid B, and LCNEC clusters), supra-carcinoids (LNEN005, LNEN012, S01513), LCNEC, and SCLC. Data necessary to reproduce the figure are provided in Supplementary Data 1, and in the European Genome-phenome Archive.



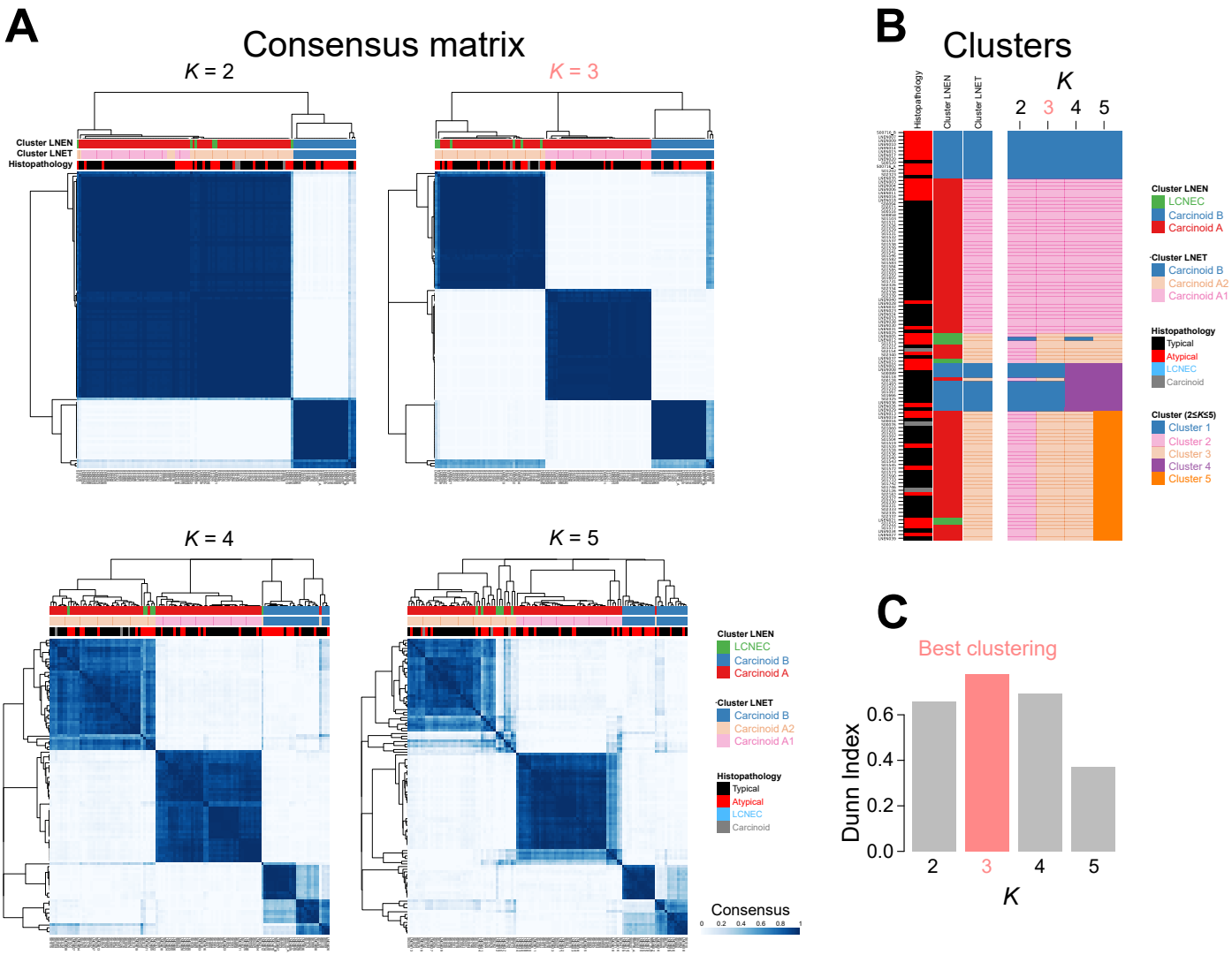
Supplementary Figure 15 Estimation of the amount of immune cells in the different pulmonary carcinoid groups from transcriptome data. The upper panel represents immune cells of each LNEN cluster and supra-carcinoids (supra-ca). The average proportion of each cell type in each group is represented. The lower panel represents the linear permutation test significance (q-value; colours: dark for q <0.001, intermediate for q <0.01, light for q <0.05, white for q ≥0.05) of the difference in cell type composition, for each cell type (row), and each possible pairwise comparison between groups (columns). Comparisons with a cell proportion difference greater than 2% are indicated by a black box. Estimates are computed using software quanTIseq (see Online methods). Data necessary to reproduce the figure are provided in Supplementary Data 1.



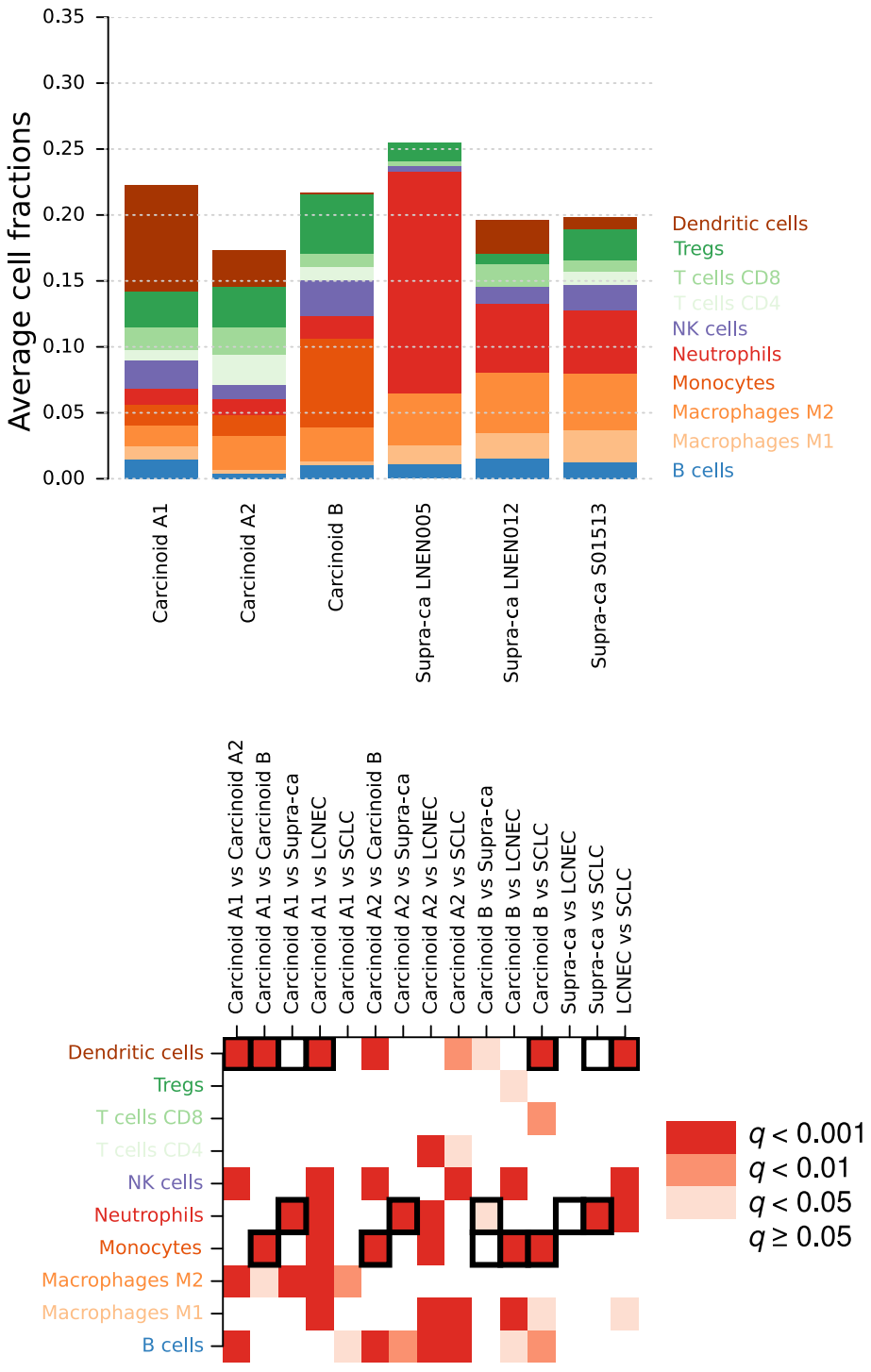
Supplementary Figure 16 Cancer-relevant somatically altered pathways altered in typical and atypical carcinoids. Colours correspond to the different types of genomic alterations. Data necessary to reproduce the figure are provided in Supplementary Data 4.



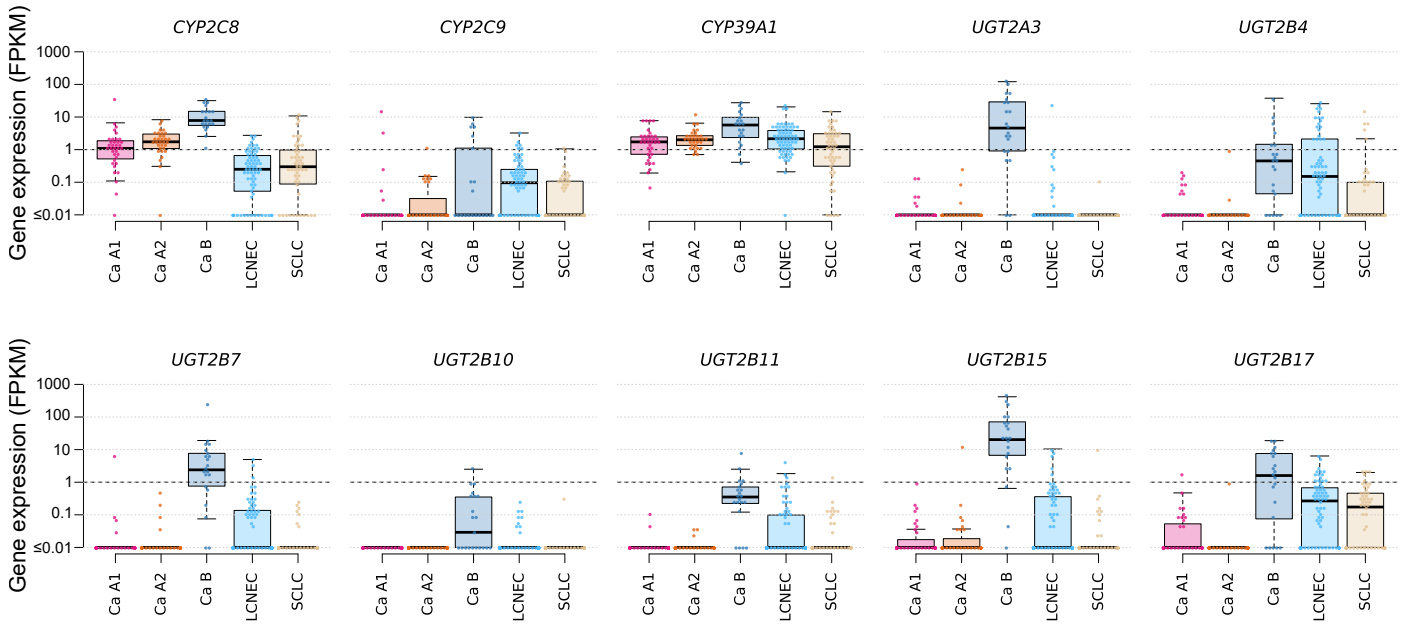
Supplementary Figure 17 Robustness of the MOFA latent factors presented in Figure 4A. A) Correlation between LF across runs for MOFA run on all LNET samples (the best run among the 20 is presented Figure 4A and Supplementary Figure 13D). B) Correlation between LF across runs for MOFA run on all LNET or SCLC samples (the best run among the 20 is presented Supplementary Figure 13C). Figure design follows that of Supplementary Figure 2. Data necessary to reproduce the figure are provided in Supplementary Data 1.



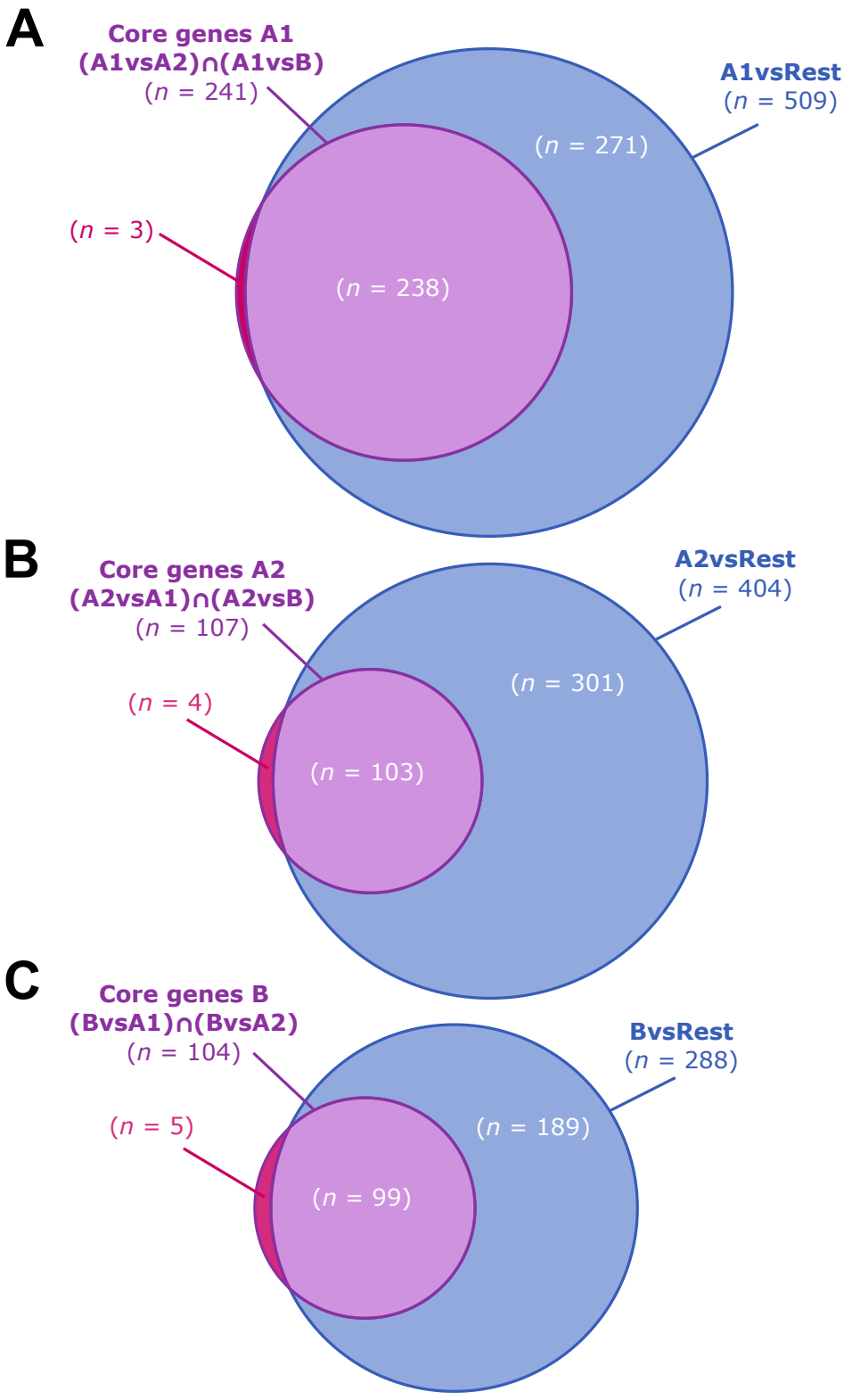
Supplementary Figure 18 Robustness of the consensus clustering of pulmonary carcinoids presented in Figure 4A. A) Heatmap of the consensus matrix for four numbers of clusters *K*; cluster memberships and histopathological types are reported above the columns, and the dendrogram represents a hierarchical clustering. B) Cluster membership as a function of *K*. C) Clustering quality metric (Dunn Index) for each value of *K*; the best clustering according to the metric is highlighted in pink. Data necessary to reproduce the figure are provided in Supplementary Data 1.



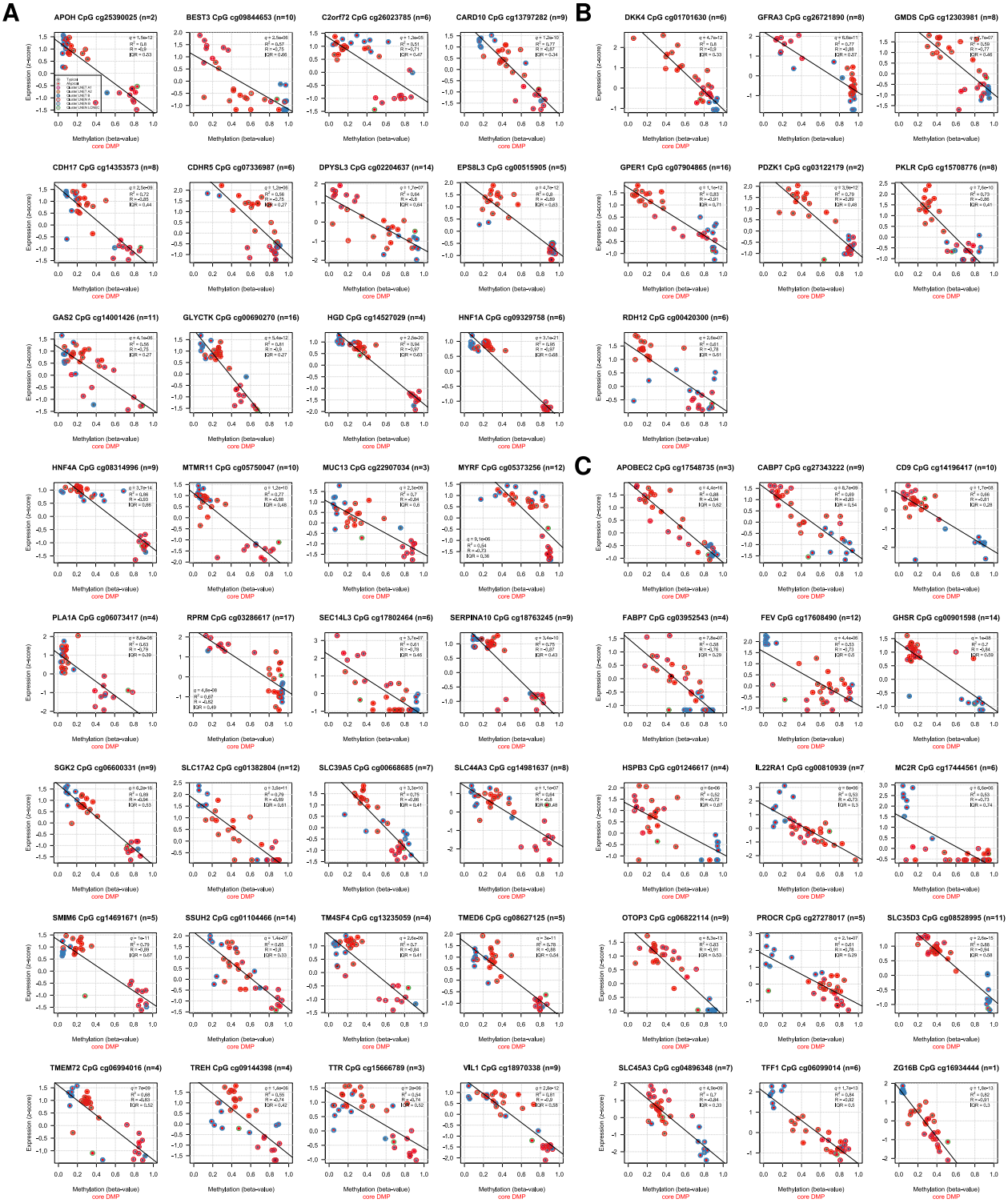
Supplementary Figure 19 Estimation of the amount of immune cells in the different LNET clusters and supra-carcinoids from transcriptome data. Figure design follows that of Supplementary Figure 15. Data necessary to reproduce the figure are provided in Supplementary Data 1.



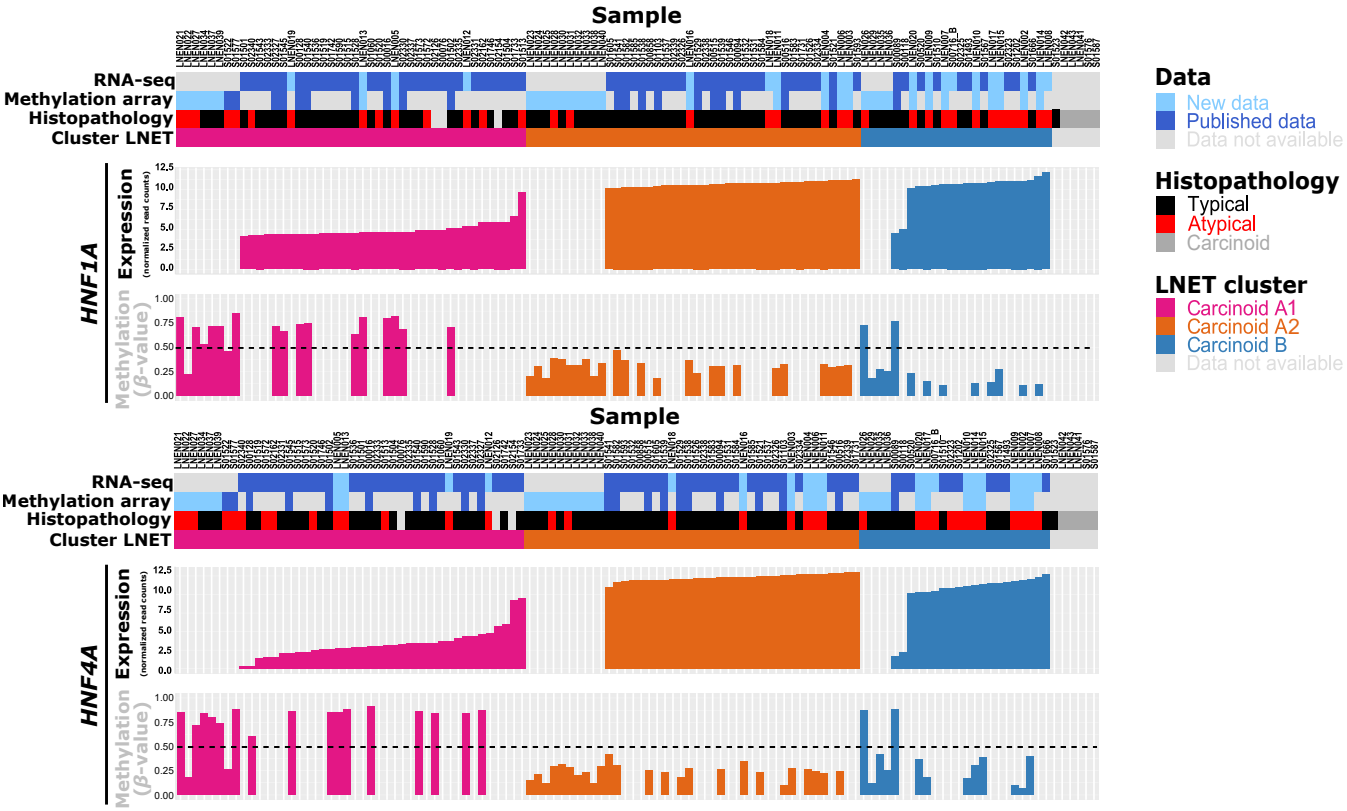
Supplementary Figure 20 Expression levels of genes involved in phase I and phase II (cytochrome P450) xenobiotic metabolism in the different LNET clusters, LCNEC and SCLC. Expression is measured in fragments per kilobase million (FPKM) units; in each plot, beeswarm plots are superimposed to boxplots to display the distribution of expression level in the corresponding groups. Centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. Data necessary to reproduce the figure are provided in Supplementary Data 1, and in the European Genome-phenome Archive.



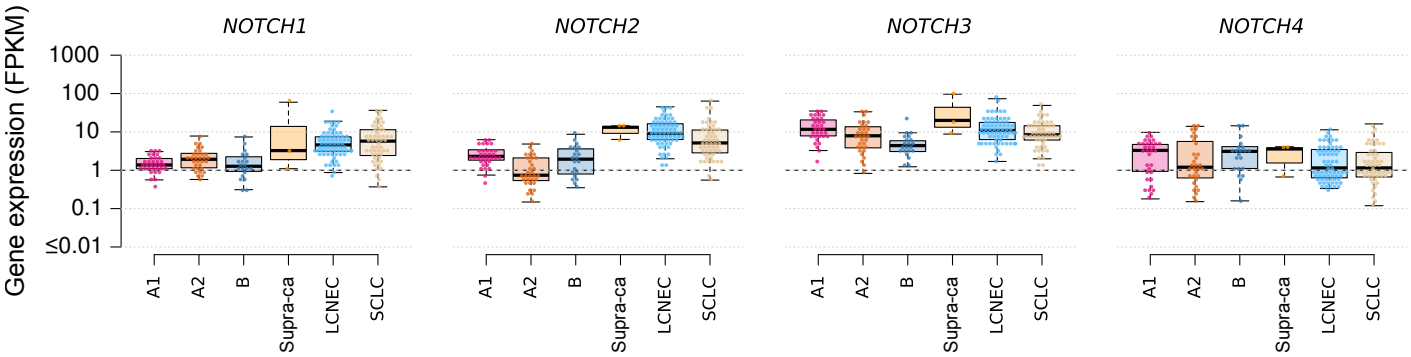
Supplementary Figure 21 Comparison of two methods to identify core differentially expressed (DE) genes of LNET clusters. Panels (A), (B), and (C) present VENN diagrams contrasting the sets of genes that are DE in all pairwise comparisons between the focal group and other groups [e.g., denoted (A1vsA2)∩(A1vsB)], and the set of genes that are DE between the focal group and all the rest (e.g., denoted A1vsRest).



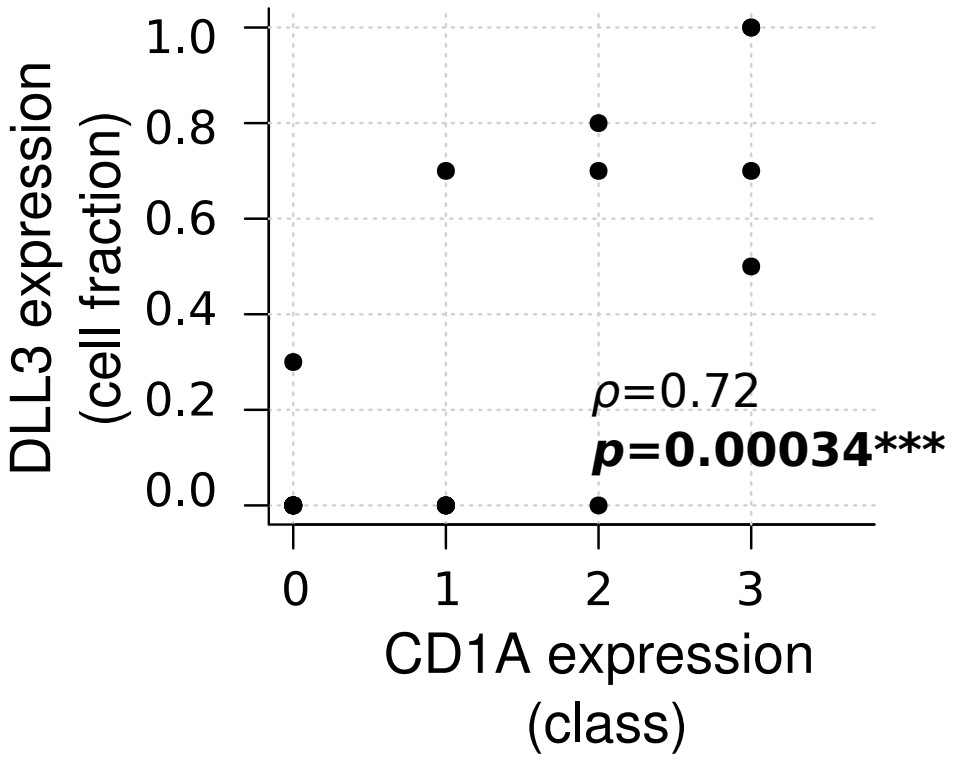
Supplementary Figure 22 Correlations between DNA methylation and gene expression for core genes of LNET clusters. Panels (A), (B), and (C) provide DNA methylation and gene expression correlations in cluster A1, A2 and B, respectively. For each coding gene, we only represent the CpGs from the promoter region and that display the strongest association (see Online Methods). Each plot represents the correlation between the β-values of the CpG and the z-scores of the corresponding gene; lines represent the best linear model fit; point colors represent the histopathological type; inner circles represent LNET clusters, outer circles represent LNEN clusters. Pearson correlation coefficients (R), corresponding correlation test q-values, and inter-quartile ranges (IQR) of the distribution of β-values of the CpG are mentioned in the top right. The number of CpGs associated with each gene, denoted by n, is mentioned in the title of each plot. If the represented CpG belongs to the core DMP of the cluster, this is mentioned in red under each plot. Data necessary to reproduce the figure are provided in Supplementary Data 10 and 11.



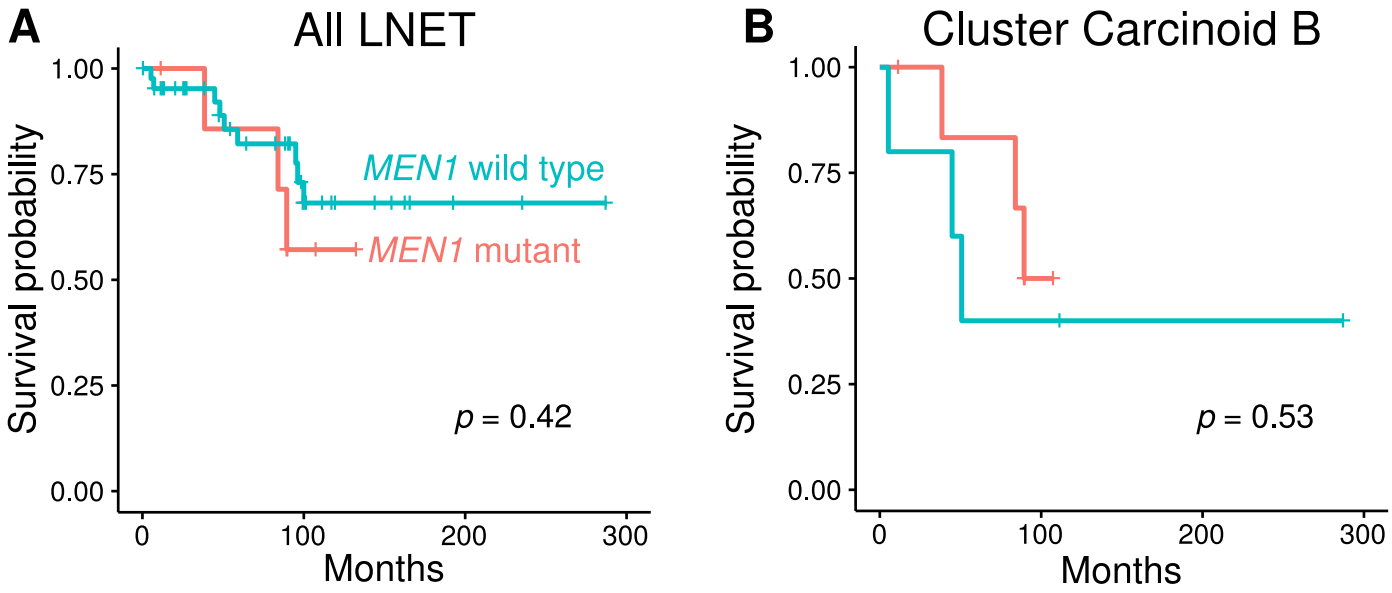
Supplementary Figure 23 DNA methylation and gene expression levels of *HNF1A* and *HNF4A* in LNET samples. DNA methylation levels correspond to the mean β-value of the CpGs correlated to the gene expression from Supplementary Data 10. Data necessary to reproduce the figure are provided in Supplementary Data 1, 10, and in the European Genome-phenome Archive.



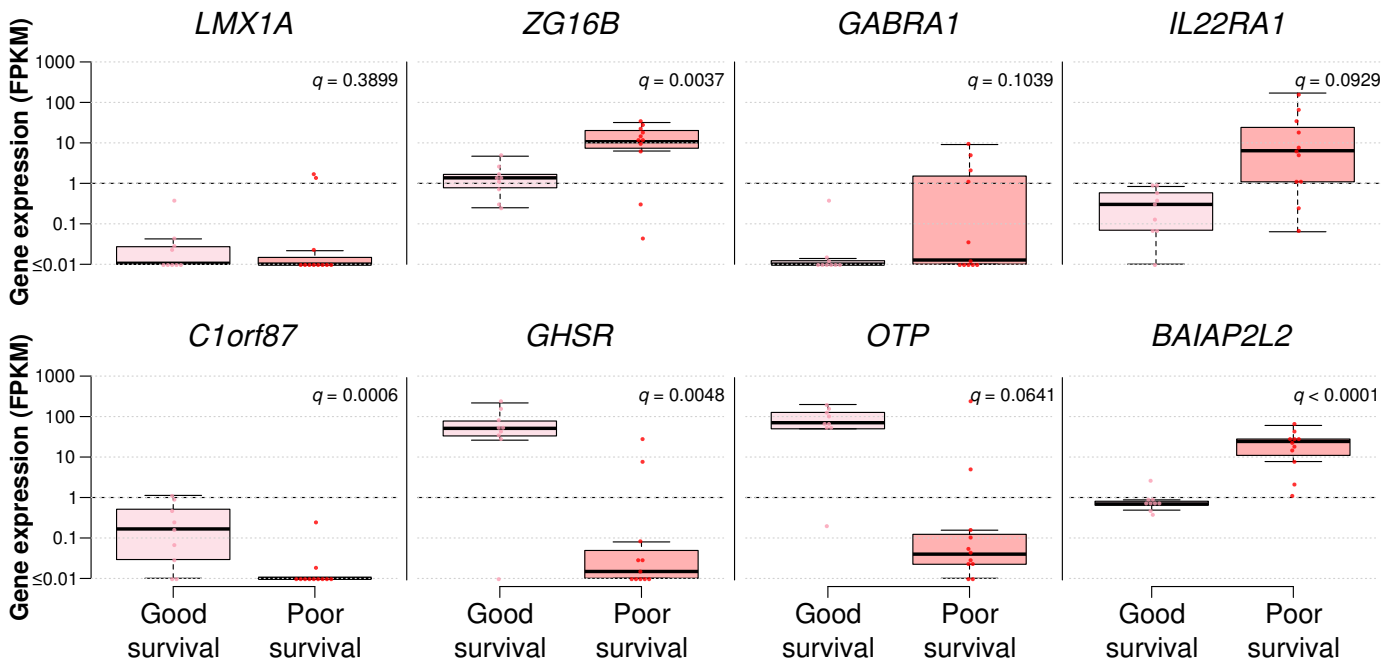
Supplementary Figure 24 Expression levels of NOTCH genes in the different LNET clusters, supra-ca, LCNEC and SCLC. The design of each panel follows that of Supplementary Figure 20. Data necessary to reproduce the figure are provided in Supplementary Data 1 and in the European Genome-phenome Archive.



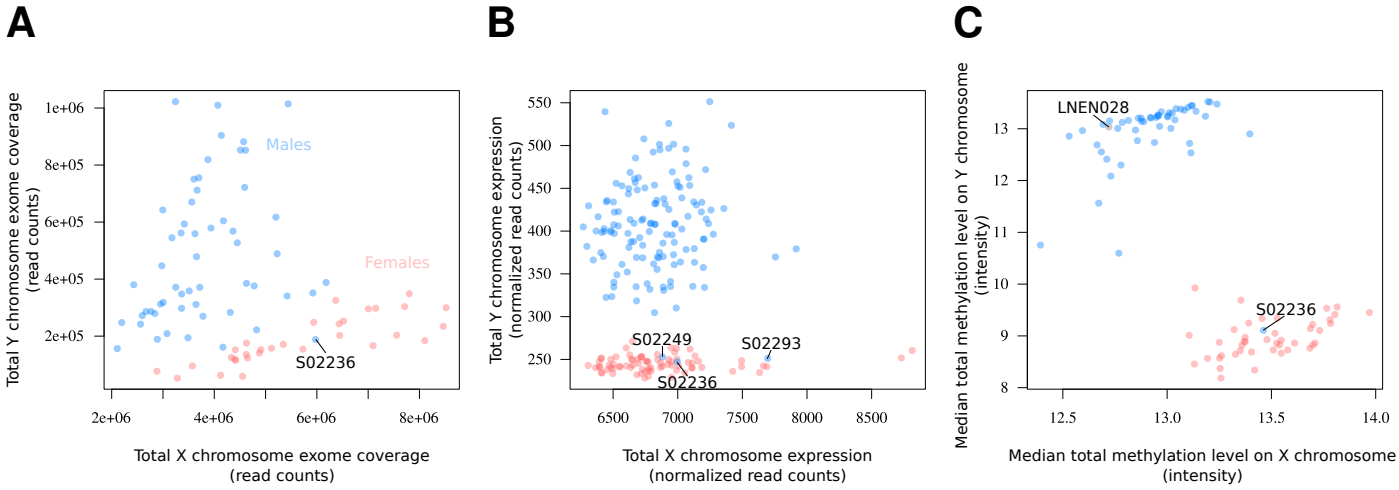
Supplementary Figure 25 Correlation between DLL3 and CDA1 expression based on immunohistochemistry in a validation series. The fraction of tumor cells exhibiting a cytoplasmic staining for DLL3 are represented on the y axis. The x axis corresponds to the CDA1 positivity classes based on the percentage of the total surface of the tumour exhibiting a membrane staining: 1 corresponds to less than 1%, 2 to a percentage between 1% and 5%, and 3 to more than 5%. The p-value and correlation coefficients of the Spearman correlation test are mentioned; $0.01 \leq p < 0.05$, $0.001 \leq p < 0.01$, and $p < 0.001$ are annotated by one, two, and three stars, respectively. Data necessary to reproduce the figure are provided in Supplementary Data 9.



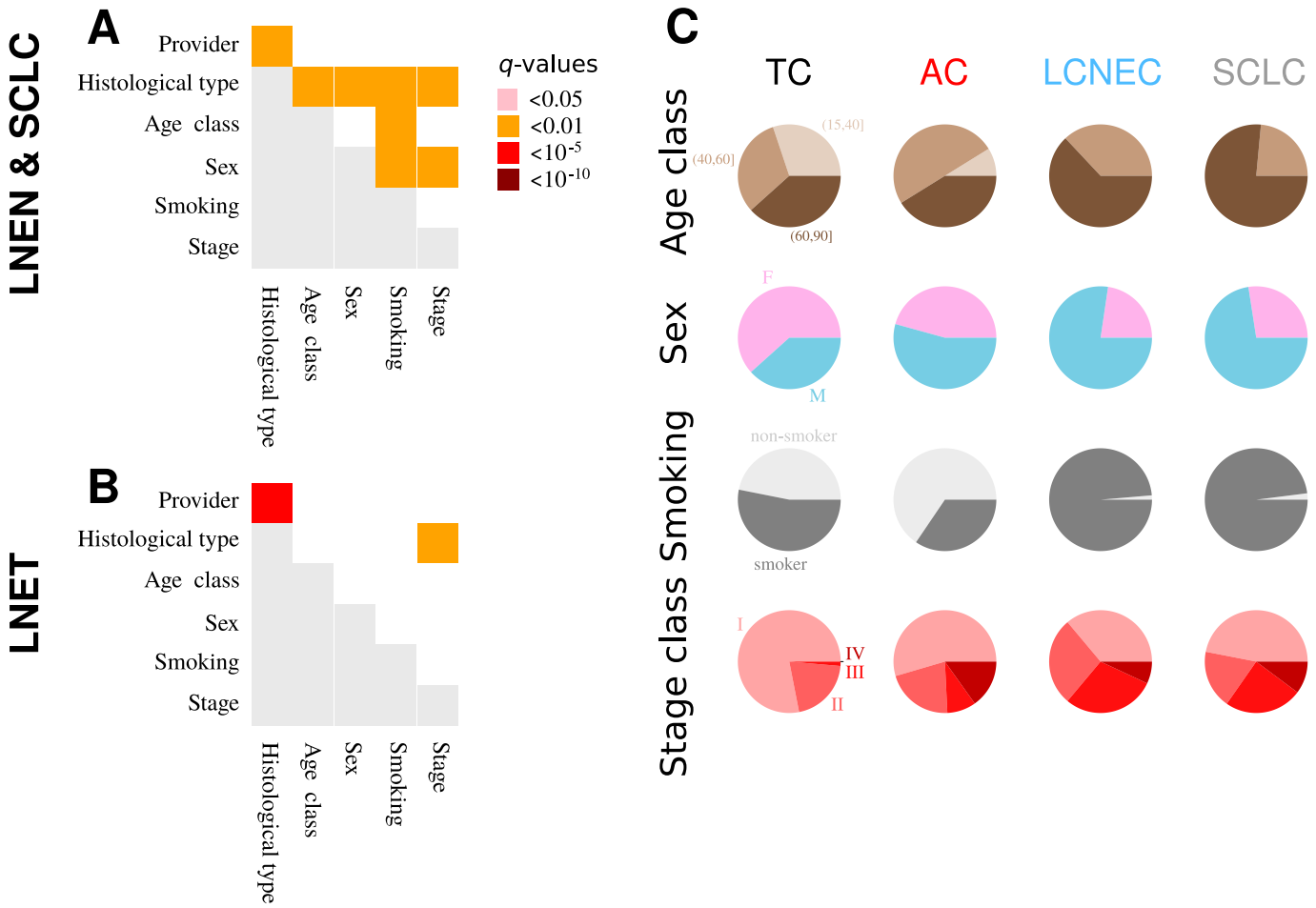
Supplementary Figure 26 Survival (Kaplan-Meier curve) of *MEN1* wild type compared to mutant cases. A) Analysis with all LNET samples. B) Analysis restricted to cluster Carcinoid B samples. The logrank test *p*-value is given at the bottom right for each panel. Data necessary to reproduce the figure are provided in Supplementary Data 1 and 4.



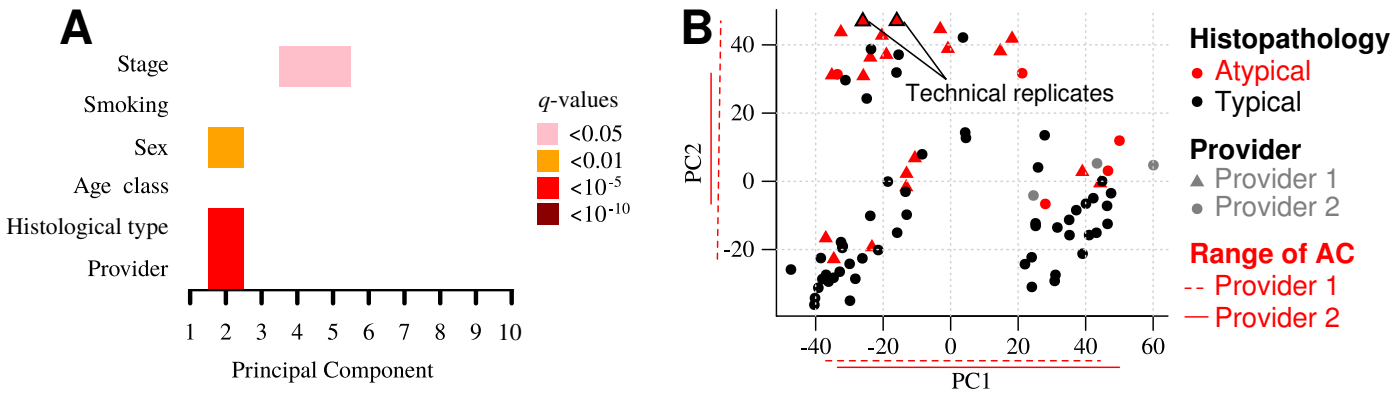
Supplementary Figure 27 Expression levels of core cluster B genes associated with survival (Figure 1B). For each gene selected by the penalized Cox regression (Supplementary Data 13), the expression levels between the good- (histopathological (HP) atypical predicted by the machine learning (ML) as typical, in pink) and poor-prognosis groups of atypical carcinoids (HP-atypical predicted as ML-atypical, in red) are compared. Expression is measured in fragments per kilobase million (FPKM) units; in each plot, beeswarm plots are superimposed to boxplots to display the distribution of expression level in the corresponding groups. Centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than the 1.5-fold IQR. The q-values corresponds to the Benjamini-Hochberg adjusted p-value of permutation tests. Data necessary to reproduce the figure are provided in Supplementary Data 1, 13, and in the European Genome-phenome Archive.



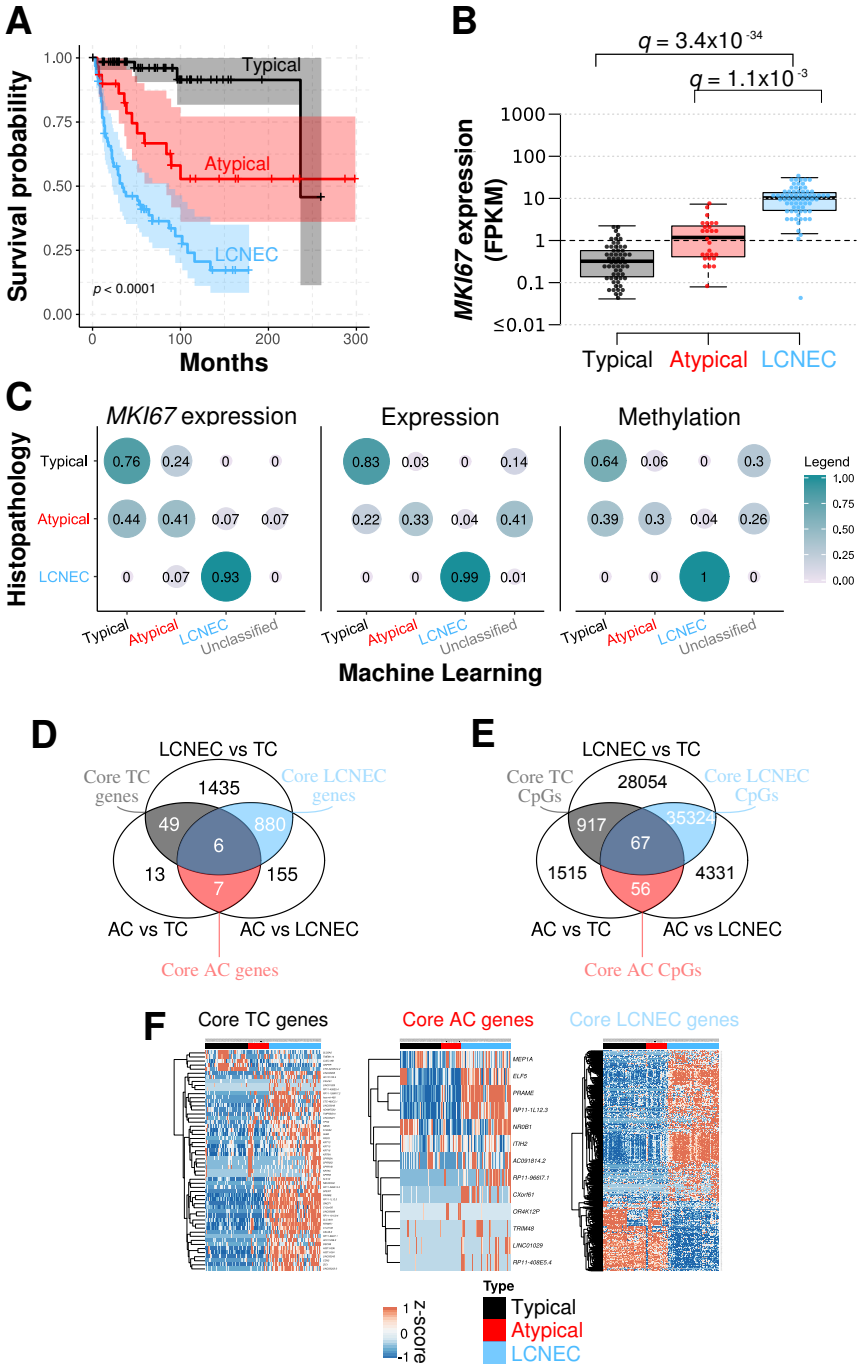
Supplementary Figure 28 Sex reclassification and multi-omic validation of reported clinical sex. A) Total exome reads coverage on the X and Y chromosomes for each sample. B) Total expression level of each sample on the X and Y chromosomes (in variance-stabilized read counts). C) Median methylation array total intensity on the X and Y chromosomes. In each panel, point colors correspond to the sexes (blue for male, red for female), and samples with discordant reported clinical sex and molecular patterns on sex chromosomes are indicated. Data necessary to reproduce the figure are provided in Supplementary Data 1, and in the European Genome-phenome Archive.



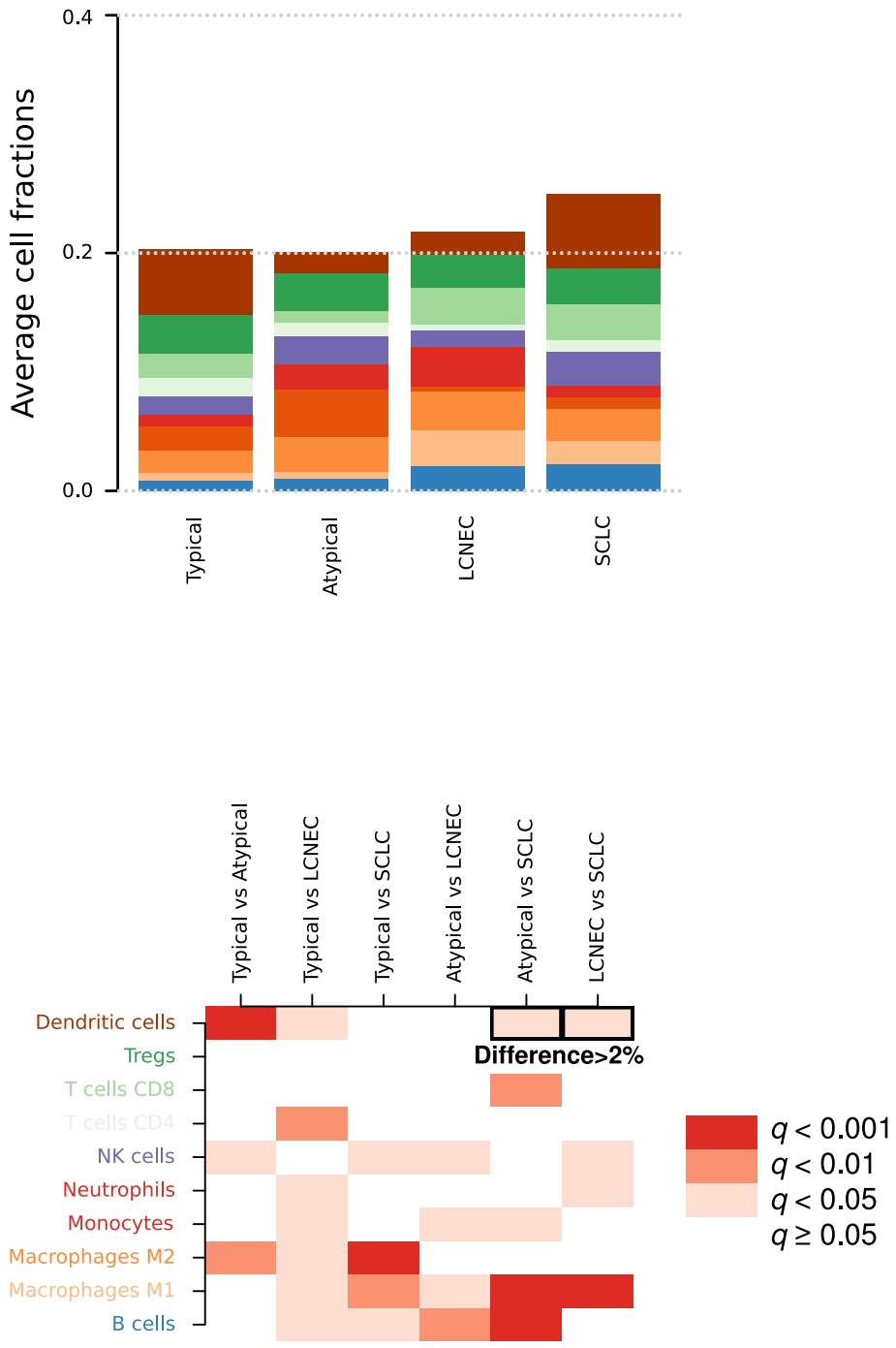
Supplementary Figure 29 Associations between clinical variables. A) Matrix of the significance (q-value) of the associations between pairs of variables, for all 242 samples from Supplementary Data 1. B) Matrix of the significance (q-value) of the association between pairs of variables, for all 116 LNET samples from Supplementary Data 1. C) Proportion of each level of each variable (rows) for each histopathological type (columns). In (A) and (B), associations are computed using Fishers exact test, adjusting for multiple testing using the Benjamini-Hochberg procedure; because of symmetry, only the upper diagonal was tested and represented. Data necessary to reproduce the figure are provided in Supplementary Data 1.



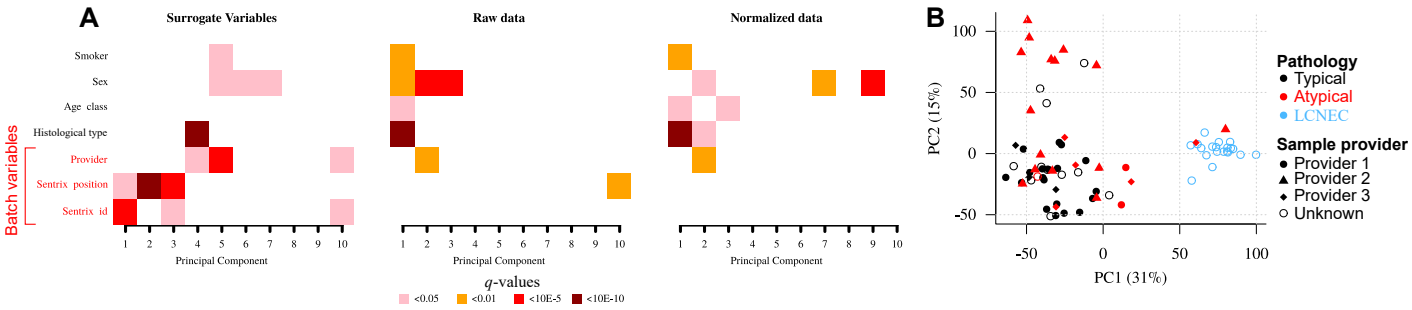
Supplementary Figure 30 Associations between clinical variables and expression profiles of LNET. A) Matrix of the significance (q-value) of the associations, computed using Fishers exact test, between clinical variables and expression principal components. B) First two axes of the PCA from panel A, with sample providers highlighted (point shapes); red segments next to the axes indicate the range of the distribution of atypical carcinoids (AC) from each provider on each principal component. Figure design follows that of Supplementary Figure 29. Data necessary to reproduce the figure are provided in Supplementary Data 1.



Supplementary Figure 31 Supervised analysis of histological types. A) Kaplan-Meier curve of overall survival of histopathological types (logrank test *p*-value is given bottom left). B) Boxplot of the expression level (in Fragments Per Kilobase Million; FPKM) of *MKI67* for each histopathological type. Centre line represents the median and box bounds represent the inter-quartile range (IQR). The whiskers span a 1.5-fold IQR or the highest and lowest observation values if they extend no further than 1.5-fold IQR. The differential expression analysis *q*-value obtained from transcriptome-wide comparisons (Supplementary Data 15) is given above each comparison. C) Machine learning analysis associated with the classification of typical carcinoids, atypical carcinoids, and LCNEC. Left panel: confusion matrix associated with the classification based on *MKI67* expression only. Middle panel: confusion matrix associated with the classification based on expression data. Right panel: confusion matrix associated with the classification based on methylation data. D) Venn diagram of core differentially expressed genes in pairwise comparisons between histopathological types. E) Venn diagram of core CpGs in pairwise comparisons between histopathological types. F) Expression of core differentially expressed genes for each histopathological type. Data necessary to reproduce the figure are provided in Supplementary Data 1, 15, and in the European Genome-phenome Archive.



Supplementary Figure 32 Estimation of the amount of immune cells in the different histopathological types from transcriptome data. Figure design follows that of Supplementary Figures 15 and 19. Data necessary to reproduce the figure are provided in Supplementary Data 1.



Supplementary Figure 33 Assessment of the batch effects in the EPIC 850K methylation array analysis. A) Matrix of the significance (q-value) of the associations, computed using Fishers exact test, between batch and clinical variables and: i) methylation surrogate variables determined from non-negative control probes (left panel), ii) the principal components of the most variable M-values (Online Methods), before functional normalization (middle panel), iii) the principal components of the most variable M-values (Online Methods), after functional normalization (right panel). B) First two axes of the PCA from panel C, with sample providers and histolopathological types highlighted (point shapes and colors, respectively). Figure design follows that of Supplementary Figures 29 and 30. Data necessary to reproduce the figure are provided in Supplementary Data 1.

Appendix B

Appendix B

B.1 Supplementary material from chapter 3

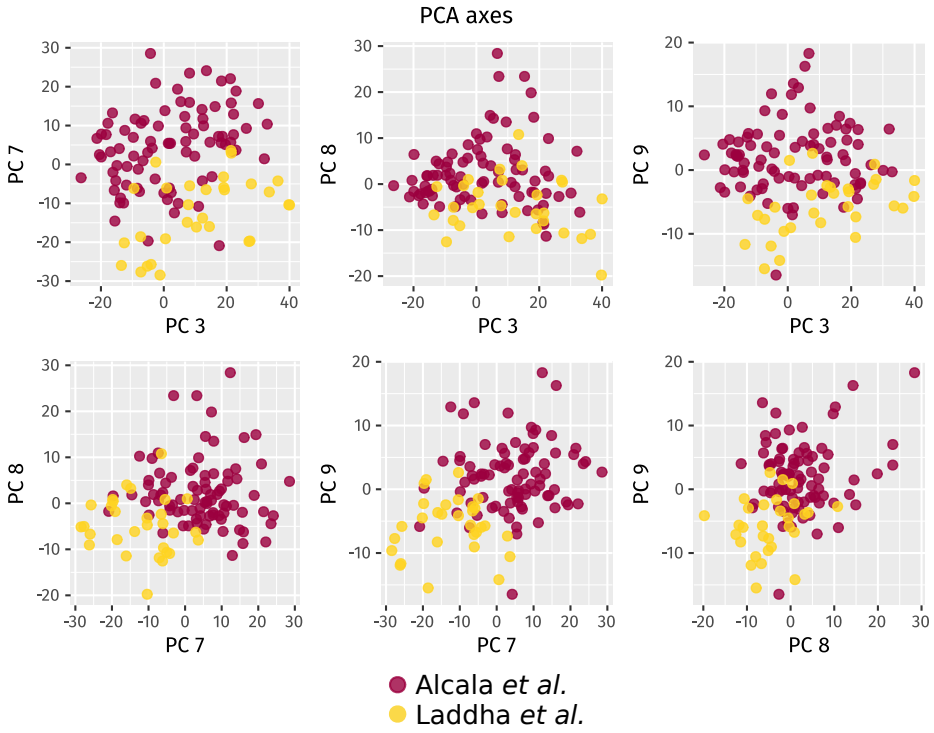


FIGURE B.1: **PCA axes correlating with study of origin**. To verify that inter-study variations were not the major sources of variations in the integrated dataset, a PCA has been performed based on the harmonized transcriptomic datasets. The figure provides pair-wise representations of the PCA axes (out of 10 axes) correlating with the study of origin.

Appendix C

Appendix C

C.1 Supplementary material from chapter 4

Imputation and samples quality control figures

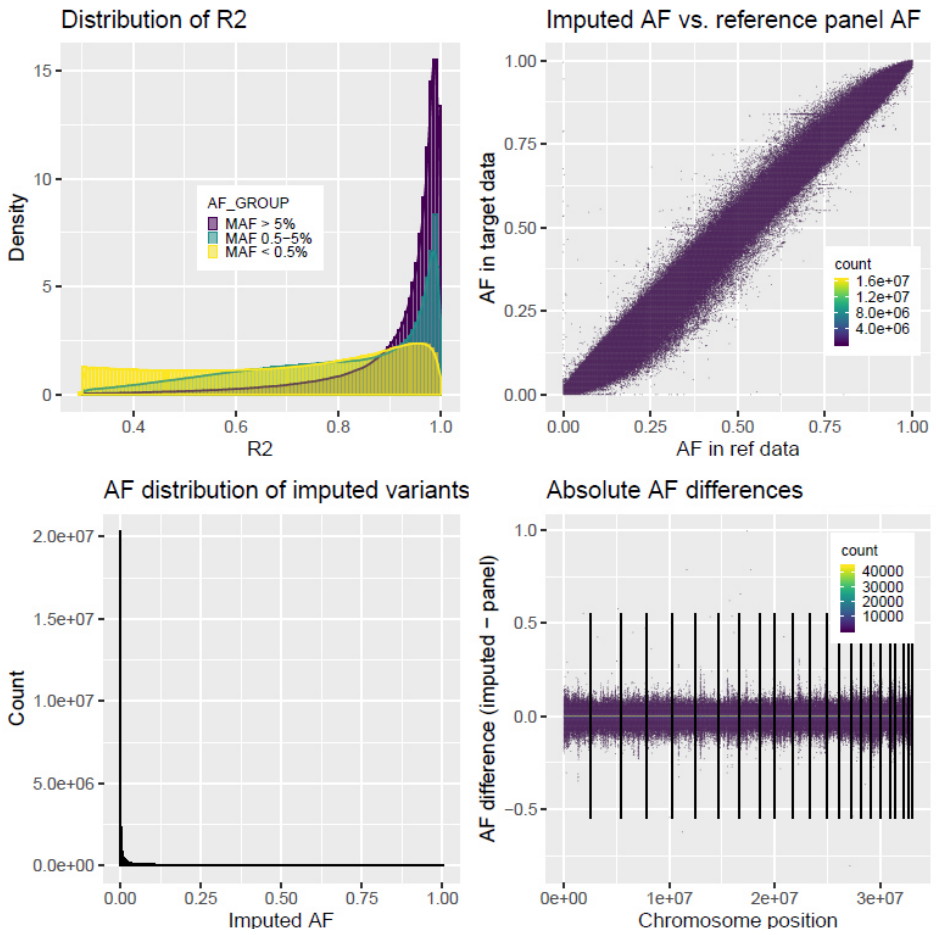


FIGURE C.1: **Imputation quality controls (European samples)**. Top left panel: minimac4 R2 quality measure distribution for each MAF category (MAF > 5%, MAF between 0.5 and 5% and MAF below 0.5%). Bottom left panel: distribution of AF in the imputed dataset. Top right panel: comparison of the SNPs AF (with an R2 value above 0.3) in the imputed data with the same SNPs AF in the 1000 Genome dataset. Bottom right panel: absolute AF difference between the imputed and 1000 Genome data across the chromosomes positions (consecutive chromosomes positions ordered on the x-axis).

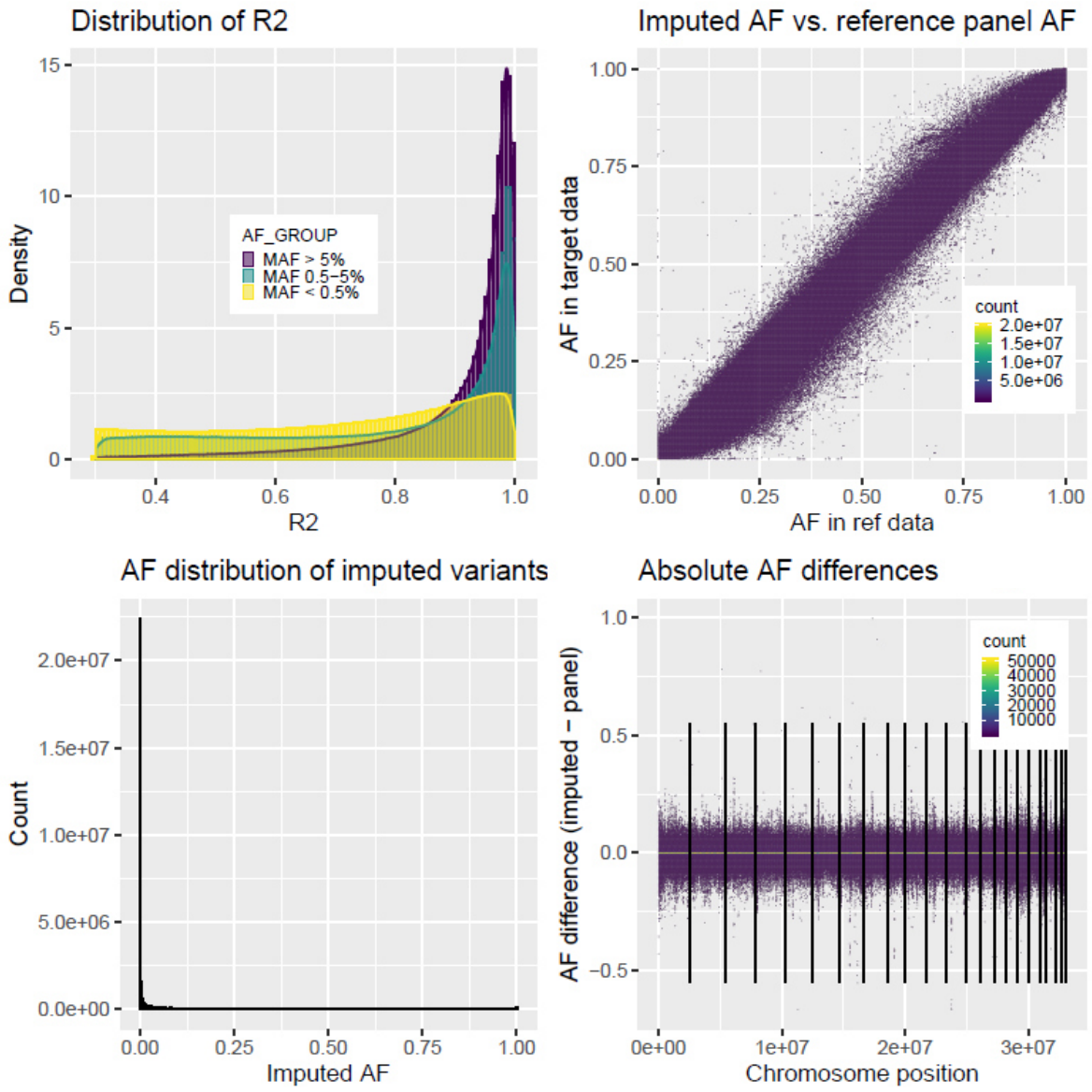


FIGURE C.2: **Imputation quality controls (Asian samples)**. Same legend as for Figure C.1

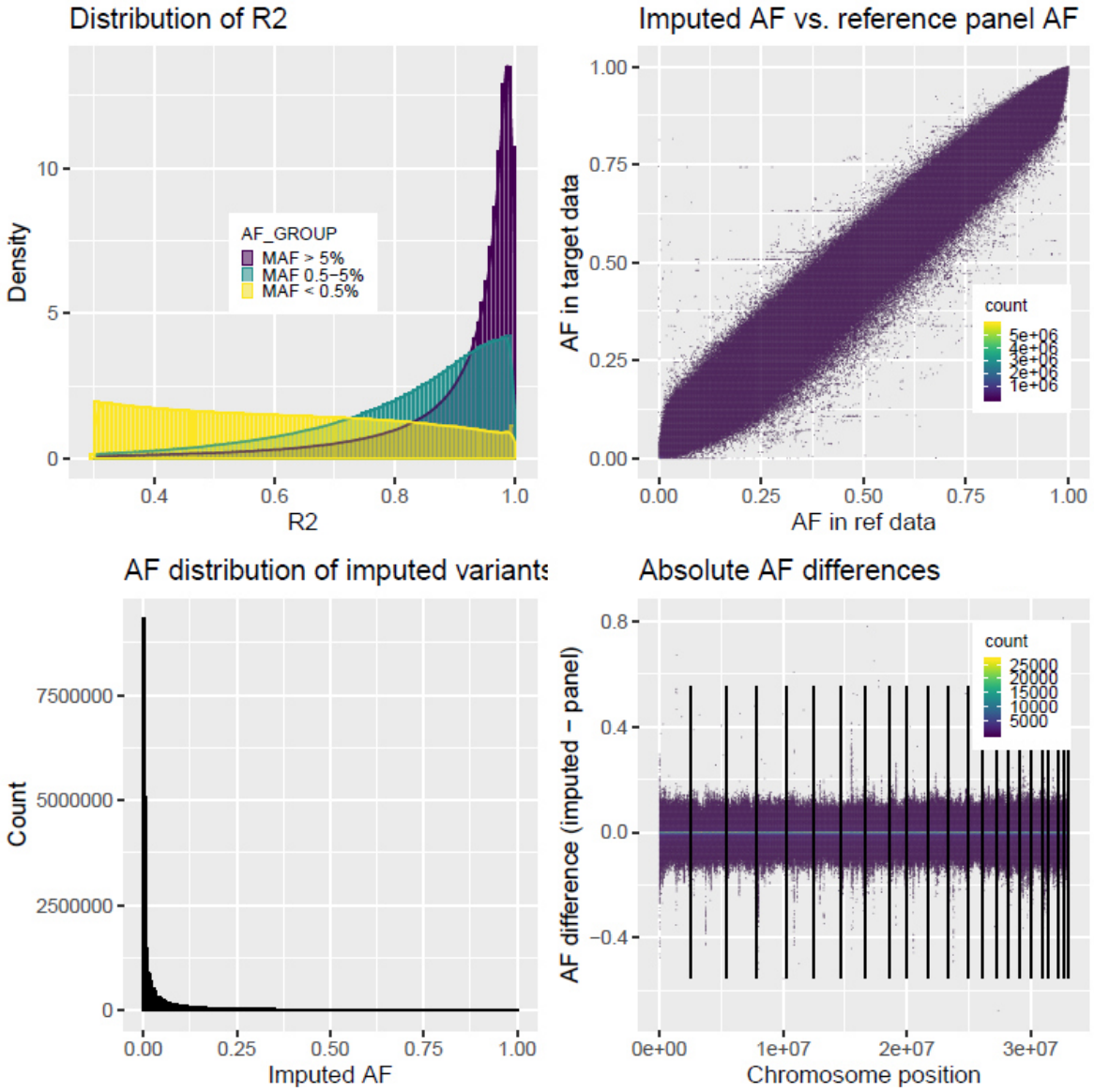


FIGURE C.3: **Imputation quality controls (African samples).** Same legend as for Figure C.1.

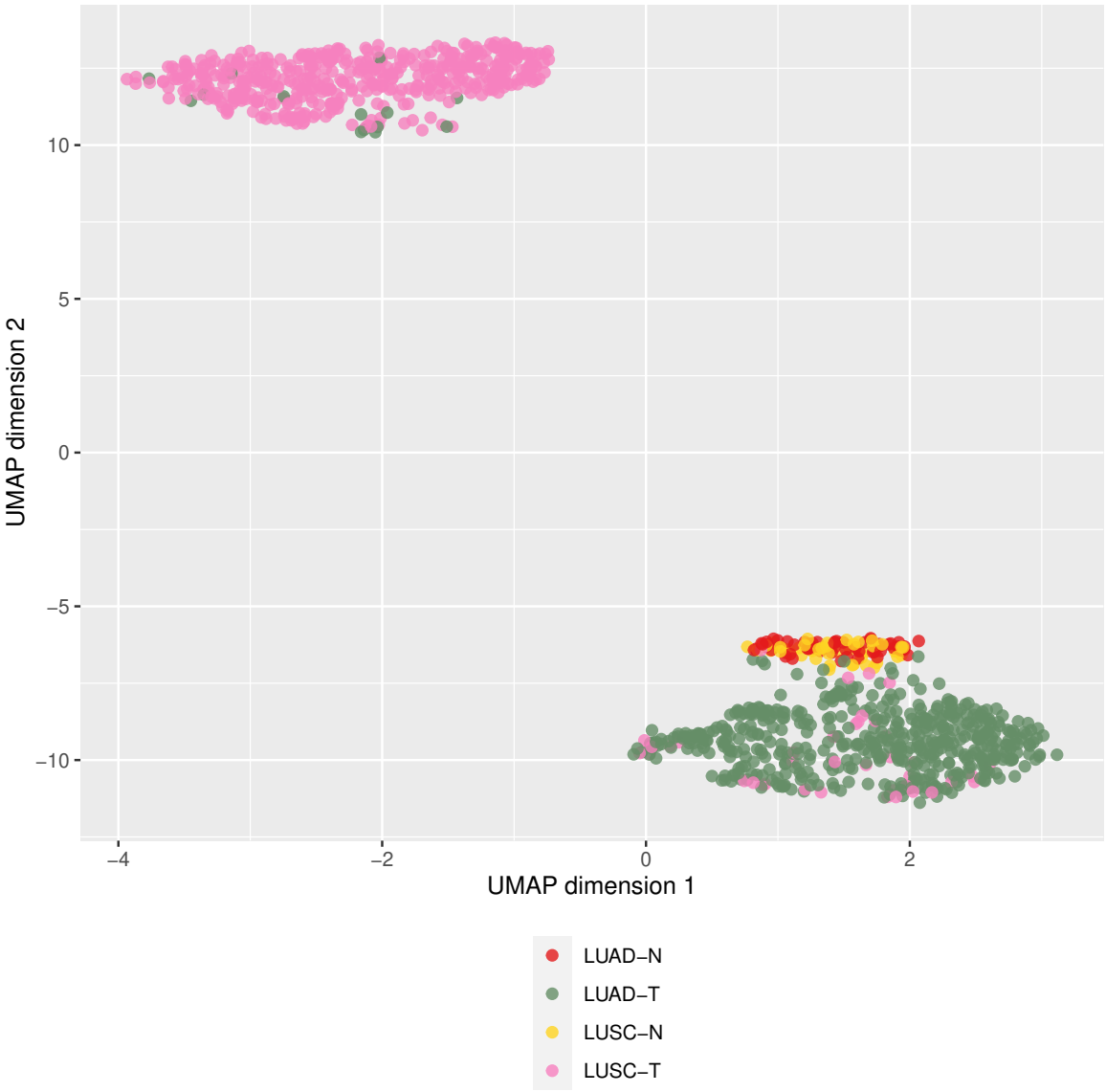


FIGURE C.4: **UMAP representation of the TCGA lung cancer samples.** The x and y-axis represent respectively the first and second dimensions resulting from UMAP dimensionality reduction. Four samples types are included in the representations: LUAD tumor and normal samples (LUAD-T and LUAD-N respectively) and LUSC tumor and normal samples (LUSC-T and LUSC-N respectively).

Bibliography

[1] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (Nov. 2018), pp. 394–424. ISSN: 1542-4863. DOI: 10.3322/caac.21492 (cit. on p. 17).

[2] Gilles R. Dagenais et al. "Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study". In: *The Lancet* 395.10226 (Mar. 2020), pp. 785–794. ISSN: 1474547X. DOI: 10.1016/S0140-6736(19)32007-0 (cit. on p. 17).

[3] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (Apr. 2009), pp. 719–724. ISSN: 00280836. DOI: 10.1038/nature07943 (cit. on pp. 17, 22–24).

[4] David S. Wishart. "Is Cancer a Genetic Disease or a Metabolic Disease?" In: *EBioMedicine* 2.6 (June 2015), pp. 478–479. ISSN: 23523964. DOI: 10.1016/j.ebiom.2015.05.022 (cit. on p. 17).

[5] Julia Eggert. "Biology of cancer". In: *Cancer Basics (Second Edition)*. Oncology Nursing Society, 2017, p. 816. ISBN: 9781935864929 (cit. on pp. 17, 23).

[6] Andreas Luch. "Nature and nurture - Lessons from chemical carcinogenesis". In: *Nature Reviews Cancer* 5.2 (Feb. 2005), pp. 113–125. ISSN: 1474175X. DOI: 10.1038/nrc1546 (cit. on p. 17).

[7] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation." In: *Cell* 144.5 (Mar. 2011), pp. 646–74. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.02.013 (cit. on pp. 17, 18, 136).

[8] Oswald T. Avery, Colin M. Macleod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III". In: *Journal of Experimental Medicine* 79.2 (Feb. 1944), pp. 137–158. ISSN: 15409538. DOI: 10.1084/jem.79.2.137 (cit. on p. 18).

[9] Thomas Hunt Morgan et al. *The Mechanism of Mendelian heredity*. New York : H. Holt and company, 1915 (cit. on p. 18).

[10] J. D. Watson and F. H.C. Crick. "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738. ISSN: 00280836. DOI: 10.1038/171737a0 (cit. on p. 18).

[11] Robert A. Weinberg. *The biology of cancer (second edition)*. Ed. by Garland Science. 2014, p. 960. ISBN: 9780815345282 (cit. on pp. 20, 24, 25).

[12] Xiaotu Ma et al. "DNA methylation data analysis and its application to cancer research". In: *Epigenomics* 5.3 (2013), pp. 301–316. DOI: 10.2217/epi.13.26 (cit. on p. 20).

[13] Bert Vogelstein et al. "Cancer genome landscapes". In: *Science* 340.6127 (Mar. 2013), pp. 1546–1558. ISSN: 10959203. DOI: 10.1126/science.1235122 (cit. on p. 23).

[14] Zbyslaw Sondka et al. "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers". In: *Nature Reviews Cancer* 18.11 (Nov. 2018), pp. 696–705. ISSN: 14741768. DOI: 10.1038/s41568-018-0060-1 (cit. on p. 23).

[15] *Cancer Gene Census*. 1999. URL: https://cancer.sanger.ac.uk/census. Online. Accessed October 2020 (cit. on p. 23).

[16] E. Premkumar Reddy et al. "A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene". In: *Nature* 300.5888 (1982), pp. 149–152. ISSN: 00280836. DOI: 10.1038/300149a0 (cit. on p. 23).

[17] A. G. Knudson. "Mutation and cancer: statistical study of retinoblastoma." In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4 (1971), pp. 820–823. ISSN: 00278424. DOI: 10.1073/pnas.68.4.820 (cit. on p. 23).

[18] Francisco Martínez-Jiménez et al. "A compendium of mutational cancer driver genes". In: *Nature Reviews Cancer* (Aug. 2020), pp. 1–18. ISSN: 14741768. DOI: 10.1038/s41568-020-0290-x (cit. on p. 23).

[19] L A Loeb. "Mutator phenotype may be required for multistage carcinogenesis." In: *Cancer research* 51.12 (June 1991), pp. 3075–9. ISSN: 0008-5472 (cit. on p. 24).

[20] Cristian Tomasetti and Bert Vogelstein. "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217 (Jan. 2015), pp. 78–81. ISSN: 10959203. DOI: 10.1126/science.1260825 (cit. on p. 24).

[21] Cristian Tomasetti, Lu Li, and Bert Vogelstein. "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* (2017). DOI: doi:10.1126/science.aaf9011 (cit. on p. 24).

[22] Song Wu et al. "Substantial contribution of extrinsic risk factors to cancer development". In: *Nature* 529 (2016), pp. 43–47. DOI: 10.1038/nature16166 (cit. on p. 24).

[23] *Lung Source: Globocan 2018 Number of new cases in 2018, both sexes, all ages*. 2018. URL: http://gco.iarc.fr/today. Online. Accessed October 2020 (cit. on pp. 24, 112).

[24] Julian Peto. "Cancer epidemiology in the last century and the next decade". In: *Nature* 411.6835 (May 2001), pp. 390–395. ISSN: 00280836. DOI: 10.1038/35077256 (cit. on p. 25).

[25] Richard Doll and A. Bradford Hill. "Smoking and carcinoma of the lung preliminary report". In: *British Medical Journal* 2.4682 (1950), pp. 739–748. ISSN: 00071447. DOI: 10.1136/bmj.2.4682.739 (cit. on pp. 25, 113).

[26] Lawrence A. Loeb and Curtis C. Harris. "Advances in chemical carcinogenesis: A historical review and prospective". In: *Cancer Research* 68.17 (Sept. 2008), pp. 6863–6872. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-08-2852 (cit. on p. 25).

[27] Helmut K. Seitz and Felix Stickel. "Acetaldehyde as an underestimated risk factor for cancer development: Role of genetics in ethanol metabolism". In: *Genes and Nutrition* 5.2 (June 2010), pp. 121–128. ISSN: 15558932. DOI: 10.1007/s12263-009-0154-1 (cit. on p. 25).

[28] Ludmil B. Alexandrov et al. "Signatures of mutational processes in human cancer". In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. ISSN: 14764687. DOI: 10.1038/nature12477 (cit. on pp. 25, 33, 40).

[29] Serena Nik-Zainal et al. "The genome as a record of environmental exposure". In: *Mutagenesis* 30 (2015), pp. 763–770. DOI: 10.1093/mutage/gev073 (cit. on pp. 25, 40, 113).

[30] Ludmil B. Alexandrov and Michael R. Stratton. "Mutational signatures: The patterns of somatic mutations hidden in cancer genomes". In: *Current Opinion in Genetics and Development* 24.1 (2014), pp. 52–60. ISSN: 0959437X. DOI: 10.1016/j.gde.2013.11.014 (cit. on p. 25).

[31] *COSMIC: Signatures of Mutational Processes in Human Cancer*. URL: https://cancer.sanger.ac.uk/cosmic/signatures_v2.tt. Online. Accessed October 2020 (cit. on p. 26).

[32] Ludmil B. Alexandrov et al. "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793 (Feb. 2020), pp. 94–101. ISSN: 14764687. DOI: 10.1038/s41586-020-1943-3 (cit. on pp. 26, 35, 40, 138).

[33] Christopher Paul Wild. "Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology". In: *Cancer Epidemiology Biomarkers and Prevention* 14.8 (Aug. 2005), pp. 1847–1850. ISSN: 10559965. DOI: 10.1158/1055-9965.EPI-05-0456 (cit. on p. 26).

[34] Leslie Roberts. "Controversial From the Start". In: *Science* 291.5507 (Feb. 2001), pp. 1–1188. ISSN: 0036-8075. DOI: 10.1126/science.291.5507.1182a (cit. on p. 26).

[35] Thomas Laframboise. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". In: *Nucleic Acids Research* 37.13 (2009), pp. 4181–4193. DOI: 10.1093/nar/gkp552 (cit. on p. 28).

[36] Chuanhua Xing et al. "Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases". In: *European Journal of Human Genetics* 24.7 (July 2016), pp. 1029–1034. ISSN: 14765438. DOI: 10.1038/ejhg.2015.244 (cit. on p. 28).

[37] Rameen Beroukhim et al. "Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays". In: *PLoS Computational Biology* 2.5 (2006), pp. 323–332. ISSN: 15537358. DOI: 10.1371/journal.pcbi.0020041 (cit. on p. 28).

[38] Amit Dutt and Rameen Beroukhim. "Single nucleotide polymorphism array analysis of cancer". In: *Current Opinion in Oncology* 19.1 (2007), pp. 43–49. ISSN: 10408746. DOI: 10.1097/CCO.0b013e328011a8c1 (cit. on p. 28).

[39] Xueying Mao, Bryan D. Young, and Yong-Jie Lu. "The Application of Single Nucleotide Polymorphism Microarrays in Cancer Research". In: *Current Genomics* 8.4 (July 2007), pp. 219–228. ISSN: 13892029. DOI: 10.2174/138920207781386924 (cit. on p. 28).

[40] John W. Belmont et al. "The international HapMap project". In: *Nature* 426.6968 (Dec. 2003), pp. 789–796. ISSN: 00280836. DOI: 10.1038/nature02168 (cit. on p. 29).

[41] Kara Rogers. *International HapMap Project*. URL: https://www.britannica.com/event/International-HapMap-Project. Online. Accessed October 2020 (cit. on p. 29).

[42] Adi L. Tarca, Roberto Romero, and Sorin Draghici. "Analysis of microarray experiments of gene expression profiling". In: *American Journal of Obstetrics and Gynecology* 195.2 (Aug. 2006), pp. 373–388. ISSN: 00029378. DOI: 10.1016/j.ajog.2006.07.001 (cit. on p. 30).

[43] Interrogate single CpG sites. *Infinium Methylation Assay*. URL: https://emea.illumina.com/science/technology/microarray/infinium-methylation-assay.html. Online. Accessed October 2020 (cit. on pp. 30, 31).

[44] KA Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Online. Accessed October 2020 (cit. on pp. 32, 34).

[45] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 14710056. DOI: 10.1038/nrg2484 (cit. on p. 33).

[46] Terrence S. Furey. "ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions". In: *Nature Reviews Genetics* 13.12 (Dec. 2012), pp. 840–852. ISSN: 14710056. DOI: 10.1038/nrg3306 (cit. on p. 33).

[47] Feng Yan et al. "From reads to insight: A hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1 (Feb. 2020), pp. 1–16. ISSN: 1474760X. DOI: 10.1186/s13059-020-1929-3 (cit. on p. 33).

[48] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50.8 (Aug. 2018), p. 96. ISSN: 20926413. DOI: 10.1038/s12276-018-0071-8 (cit. on p. 34).

[49] Francesca Finotello et al. "Next-generation computational tools for interrogating cancer immunity". In: *Nature Reviews Genetics* 20.12 (Dec. 2019), pp. 724–746. ISSN: 14710064. DOI: 10.1038/s41576-019-0166-7 (cit. on p. 34).

[50] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge". In: *Wspolczesna Onkologia, Contemporary oncology* 19.1A (2015), A68–A77. ISSN: 14282526. DOI: 10.5114/wo.2014.47136 (cit. on p. 34).

[51] John N. Weinstein et al. "The cancer genome atlas pan-cancer analysis project". In: *Nature Genetics* 45.10 (Oct. 2013), pp. 1113–1120. ISSN: 15461718. DOI: 10.1038/ng.2764 (cit. on pp. 35, 36, 87).

[52] NCI Genomic Data Commons. *PanCanAtlas Publications*. URL: https://gdc.cancer.gov/about-data/publications/pancanatlas. Online. Accessed October 2020 (cit. on pp. 35, 124).

[53] Thomas J. Hudson et al. "International network of cancer genome projects". In: *Nature* 464.7291 (Apr. 2010), pp. 993–998. ISSN: 00280836. DOI: 10.1038/nature08987 (cit. on pp. 35, 36).

[54] Marcin Cieslik and Arul M. Chinnaiyan. "Global genomics project unravels cancer's complexity at unprecedented scale". In: *Nature* 578.7793 (Feb. 2020), pp. 39–40. ISSN: 14764687. DOI: 10.1038/d41586-020-00213-2 (cit. on pp. 35, 143, 148).

[55] Peter J. Campbell et al. "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793 (Feb. 2020), pp. 82–93. ISSN: 14764687. DOI: 10.1038/s41586-020-1969-6 (cit. on pp. 35, 137).

[56] Yilong Li et al. "Patterns of somatic structural variation in human cancer genomes". In: *Nature* 578.7793 (Feb. 2020), pp. 112–121. ISSN: 14764687. DOI: 10.1038/s41586-019-1913-9 (cit. on p. 35).

[57] Orli G. Bahcall. "UK Biobank — a new era in genomic medicine". In: *Nature Reviews Genetics* 19.12 (Dec. 2018), p. 737. ISSN: 14710064. DOI: 10.1038/s41576-018-0065-3 (cit. on p. 36).

[58] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (Oct. 2018), pp. 203–209. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z (cit. on p. 36).

[59] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1005–D1012. ISSN: 13624962. DOI: 10.1093/nar/gky1120 (cit. on p. 36).

[60] *TCGA Timeline Milestones was originally published by the National Cancer Institute.* URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline. Online. Accessed October 2020 (cit. on p. 37).

[61] Mark A. Jensen et al. "The NCI Genomic Data Commons as an engine for precision medicine". In: *Blood* 130.4 (July 2017), pp. 453–459. ISSN: 15280020. DOI: 10.1182/blood-2017-03-735654 (cit. on pp. 37, 99).

[62] Galen F. Gao et al. "Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data". In: *Cell Systems* 9.1 (July 2019), pp. 24–34. ISSN: 2405-4712. DOI: 10.1016/J.CELS.2019.06.006 (cit. on pp. 37, 88).

[63] NCI Genomic Data Commons. *PanCanAtlas Publications*. URL: https://gdc.cancer.gov/about-data/publications/pancanatlas. Online. Accessed October 2020 (cit. on p. 37).

[64] Vésteinn Thorsson et al. "The Immune Landscape of Cancer." In: *Immunity* 48.4 (Apr. 2018), pp. 812–830. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2018.03.023 (cit. on pp. 37, 136).

[65] Theo A. Knijnenburg et al. "Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas". In: *Cell Reports* 23.1 (2018), pp. 239–254. ISSN: 22111247. DOI: 10.1016/j.celrep.2018.03.076 (cit. on p. 37).

[66] Jessica W. Lau et al. "The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research". In: *Cancer Research* 77.21 (Nov. 2017), e3–e6. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-0387 (cit. on p. 37).

[67] Sheila M. Reynolds et al. "The ISB cancer genomics cloud: A flexible cloud-based platform for cancer genomics research". In: *Cancer Research* 77.21 (Nov. 2017), e7–e10. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-0617 (cit. on p. 37).

[68] Sergei Yakneen et al. "Butler enables rapid cloud-based analysis of thousands of human genomes". In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 288–292. ISSN: 15461696. DOI: 10.1038/s41587-019-0360-3 (cit. on p. 37).

[69] W.D. Travis. "Advances in neuroendocrine lung tumors". In: *Annals of Oncology* 21 (Oct. 2010), pp. vii65–vii71. ISSN: 09237534. DOI: 10.1093/annonc/mdq380 (cit. on p. 38).

[70] Jules L. Derks et al. "New Insights into the Molecular Characteristics of Pulmonary Carcinoids and Large Cell Neuroendocrine Carcinomas, and the Impact on Their Clinical Management". In: *Journal of Thoracic Oncology* 13.6 (June 2018), pp. 752–766. ISSN: 15561380. DOI: 10.1016/j.jtho.2018.02.002 (cit. on pp. 38, 39, 41).

[71] Michele Simbolo et al. "Exploring the molecular and biological background of lung neuroendocrine tumours." In: *Journal of thoracic disease* 11.Suppl 9 (May 2019), S1194–S1198. ISSN: 2072-1439. DOI: 10.21037/jtd.2019.03.66 (cit. on pp. 38, 57).

[72] Cancer.Net. *Lung Cancer - Non-Small Cell: Statistics*. URL: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics. Online. Accessed October 2020 (cit. on p. 38).

[73] Katerina Politi and Roy S. Herbst. "Lung cancer in the era of precision medicine". In: *Clinical Cancer Research* 21.10 (May 2015), pp. 2213–2220. ISSN: 15573265. DOI: 10.1158/1078-0432.CCR-14-2748 (cit. on pp. 38, 40, 55).

[74] William D Travis et al. "The 2015 World Health Organization Classification of Lung Tumors". In: *Journal of Thoracic Oncology* 10 (2015), pp. 1243–1260. DOI: 10.1097/JTO.0000000000000630 (cit. on p. 38).

[75] Inigo Martincorena and Peter J Campbell. "Somatic mutation in cancer and normal cells (Erratum)". In: *Science* 351.6277 (Mar. 2016), aaf5401–aaf5401. ISSN: 0036-8075. DOI: 10.1126/science.aaf5401 (cit. on pp. 39, 87).

[76] Viviane Teixeira Loiola de Alencar, Maria Nirvana Formiga, and Vladmir Cláudio Cordeiro de Lima. "Inherited lung cancer: A review". In: *ecancermedicalscience* 14 (Jan. 2020). ISSN: 17546605. DOI: 10.3332/ECANCER.2020.1008 (cit. on pp. 39, 112).

[77] C. I. Amos, W. Xu, and M. R. Spitz. "Is There a Genetic Basis for Lung Cancer Susceptibility?" In: *Chemoprevention of Cancer*. Vol. 151. 1999, pp. 3–12. DOI: 10.1007/978-3-642-59945-3{_}1 (cit. on p. 39).

[78] Yohan Bosse and Christopher I. Amos. "A decade of GWAS results in lung cancer". In: *Cancer Epidemiology Biomarkers and Prevention* 27.4 (Apr. 2018), pp. 363–379. ISSN: 10559965. DOI: 10.1158/1055-9965.EPI-16-0794 (cit. on pp. 39, 111, 113, 135).

[79] Thorgeir E. Thorgeirsson et al. "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease". In: *Nature* 452.7187 (Apr. 2008), pp. 638–642. ISSN: 14764687. DOI: 10.1038/nature06846 (cit. on pp. 39, 113, 135).

[80] James D McKay et al. "Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes". In: *Nature Genetics* 49.7 (July 2017), pp. 1126–1132. ISSN: 1061-4036. DOI: 10.1038/ng.3892 (cit. on pp. 39, 113, 122, 135).

[81] Michael S. Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457 (2013), pp. 214–218. ISSN: 00280836. DOI: 10.1038/nature12213 (cit. on p. 40).

[82] Kenichi Yoshida et al. "Tobacco smoking and somatic mutations in human bronchial epithelium". In: *Nature* 578.7794 (Feb. 2020), pp. 266–272. ISSN: 14764687. DOI: 10.1038/s41586-020-1961-1 (cit. on p. 40).

[83] Eric A. Collisson et al. "Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network". In: *Nature* 511.7511 (July 2014), pp. 543–550. ISSN: 14764687. DOI: 10.1038/nature13385 (cit. on p. 40).

[84] Neal I. Lindeman et al. "Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the ". In: *Journal of Thoracic Oncology* 13.3 (Mar. 2018), pp. 323–358. ISSN: 15561380. DOI: 10.1016/j.jtho.2017.12.001 (cit. on p. 40).

[85] The Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers". In: *Nature* 489.7417 (Sept. 2012), pp. 519–525. ISSN: 0028-0836. DOI: 10.1038/nature11404 (cit. on p. 40).

[86] Martin Peifer et al. "Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer". In: *Nature Genetics* 44.10 (Oct. 2012), pp. 1104–1110. ISSN: 1061-4036. DOI: 10.1038/ng.2396 (cit. on pp. 41, 57, 58).

[87] Julie George et al. "Comprehensive genomic profiles of small cell lung cancer". In: *Nature* 524.7563 (Aug. 2015), pp. 47–53. ISSN: 0028-0836. DOI: 10.1038/nature14664 (cit. on pp. 41, 57, 58, 90).

[88] Lynnette Fernandez-Cuesta and Matthieu Foll. "Molecular studies of lung neuroendocrine neoplasms uncover new concepts and entities". In: *Translational Lung Cancer Research* 8.S4 (Dec. 2019), S430–S434. ISSN: 22186751. DOI: 10.21037/tlcr.2019.11.08 (cit. on pp. 41, 56, 57).

[89] Lynnette Fernandez-Cuesta et al. "Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer". In: *EBioMedicine* 10 (Aug. 2016), pp. 117–123. ISSN: 23523964. DOI: 10.1016/j.ebiom.2016.06.032 (cit. on p. 41).

[90] Joshua D. Cohen et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test". In: *Science* 359.6378 (Feb. 2018), pp. 926–930. ISSN: 10959203. DOI: 10.1126/science.aar3247 (cit. on pp. 41, 55).

[91] Tiffany M Delhomme et al. "Needlestack: an ultra-sensitive variant caller for multi-sample next generation sequencing data". In: *NAR Genomics and Bioinformatics* 2.2 (June 2020). ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa021 (cit. on p. 41).

[92] Lynnette Fernandez-Cuesta et al. "Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids." In: *Nature communications* 5 (Mar. 2014), p. 3518. ISSN: 2041-1723. DOI: 10.1038/ncomms4518 (cit. on pp. 41, 57, 58, 90).

[93] Julie George et al. "Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors". In: *Nature Communications* 9.1 (Dec. 2018), p. 1048. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03099-x (cit. on pp. 41, 57, 58, 90, 94).

[94] Natasha Rekhtman et al. "Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma–like and Non–Small Cell Carcinoma–like Subsets". In: *Clinical Cancer Research* 22.14 (July 2016), pp. 3618–3629. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-15-2946 (cit. on pp. 41, 63).

[95] Michele Simbolo et al. "Gene Expression Profiling of Lung Atypical Carcinoids and Large Cell Neuroendocrine Carcinomas Identifies Three Transcriptomic Subtypes with Specific Genomic Alterations". In: *Journal of Thoracic Oncology* 14.9 (Sept. 2019), pp. 1651–1661. ISSN: 15560864. DOI: 10.1016/j.jtho.2019.05.003 (cit. on pp. 41, 89).

[96] Charles M. Rudin et al. "Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data". In: *Nature Reviews Cancer* 19.5 (May 2019), pp. 289–297. ISSN: 14741768. DOI: 10.1038/s41568-019-0133-9 (cit. on p. 41).

[97] Naomi Altman and Martin Krzywinski. "The curse(s) of dimensionality". In: *Nature Methods* 15.6 (June 2018), pp. 399–400. ISSN: 1548-7091. DOI: 10.1038/s41592-018-0019-x (cit. on p. 42).

[98] Pedro Domingos. "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (2012). DOI: 10.1145/2347736.2347755 (cit. on pp. 42, 46).

[99] Tom Ronan, Zhijie Qi, and Kristen M. Naegle. "Avoiding common pitfalls when clustering biological data". In: *Science Signaling* 9.432 (June 2016), re6–re6. ISSN: 19379145. DOI: 10.1126/scisignal.aad1932 (cit. on pp. 42, 43, 97, 98).

[100] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Ed. by Cambridge University Press. 2019, p. 402 (cit. on p. 43).

[101] Tom M. Mitchell. *Machine learning*. McGraw-Hill Science/Engineering/Math, 1997, p. 432 (cit. on p. 43).

[102] Maxwell W. Libbrecht and William Stafford Noble. "Machine learning applications in genetics and genomics". In: *Nature Reviews Genetics* 16.6 (June 2015), pp. 321–332. ISSN: 1471-0056. DOI: 10.1038/nrg3920 (cit. on pp. 44, 144).

[103] Danielle Denisko and Michael M. Hoffman. "Classification and interaction in random forests". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.8 (Feb. 2018), pp. 1690–1692. ISSN: 10916490. DOI: 10.1073/pnas.1800256115 (cit. on p. 45).

[104] Bradley Boehmke and Greenwell Brandon. *Hands-On Machine Learning with R*. URL: https://bradleyboehmke.github.io/HOML/process.html. Online. Accessed October 2020 (cit. on p. 47).

[105] Nikolay Oskolkov. *Unsupervised OMICs Integration*. 2019. URL: https://towardsdatascience.com/unsupervised-omics-integration-688bf8fa49bf. Online. Accessed October 2020 (cit. on p. 47).

[106] Gökcen Eraslan et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* (Apr. 2019), p. 1. ISSN: 1471-0056. DOI: 10.1038/s41576-019-0122-6 (cit. on pp. 47, 144, 145).

[107] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Ed. by Springer. 2017, p. 745 (cit. on pp. 48, 52).

[108] Genevieve L Stein-O'brien et al. "Enter the Matrix: Factorization Uncovers Knowledge from Omics Determining the Dimensions of Biology from Omics Data". In: *Trends in Genetics* 34 (2018), pp. 790–805. DOI: 10.1016/j.tig.2018.07.003 (cit. on p. 49).

[109] Ludmil B. Alexandrov et al. "Deciphering Signatures of Mutational Processes Operative in Human Cancer". In: *Cell Reports* 3.1 (Jan. 2013), pp. 246–259. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.12.008 (cit. on p. 49).

[110] Laurens Van Der Maaten and Geoffrey Hinton. *Visualizing Data using t-SNE*. Tech. rep. 2008, pp. 2579–2605 (cit. on p. 50).

[111] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *ArXiv* (Feb. 2018) (cit. on pp. 50, 51, 91, 125).

[112] Leland McInnes. *How UMAP Works — umap 0.4 documentation*. 2018. URL: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html. Online. Accessed October 2020 (cit. on pp. 50, 51).

[113] Andy Coenen and Adam Pearce. *Understanding UMAP*. URL: https://pair-code.github.io/understanding-umap/. Online. Accessed October 2020 (cit. on pp. 50, 51).

[114] Ricard Argelaguet et al. "Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets." In: *Molecular systems biology* 14.6 (June 2018), e8124. ISSN: 1744-4292. DOI: 10.15252/msb.20178124 (cit. on pp. 51, 59).

[115] Ricard Argelaguet et al. "MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1 (May 2020), p. 111. ISSN: 1474760X. DOI: 10.1186/s13059-020-02015-1 (cit. on p. 51).

[116] Kim-Ahn Lê Cao. *Webinar mixOmics: PLS methods*. 2020 (cit. on p. 52).

[117] Florian Rohart et al. "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLOS Computational Biology* 13.11 (Nov. 2017). Ed. by Dina Schneidman, e1005752. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752 (cit. on p. 52).

[118] Amrit Singh et al. "DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17 (2019), pp. 3055–3062. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty1054 (cit. on p. 52).

[119] Konrad J Karczewski and Michael P Snyder. "Integrative omics for health and disease." In: *Nature reviews. Genetics* 19.5 (May 2018), pp. 299–310. ISSN: 1471-0064. DOI: 10.1038/nrg.2018.4 (cit. on pp. 52, 56).

[120] W. D. Travis et al. "Paradigm shifts in lung cancer as defined in the new IASLC/ATS/ERS lung adenocarcinoma classification". In: *European Respiratory Journal* 38.2 (Aug. 2011), pp. 239–243. ISSN: 09031936. DOI: 10.1183/09031936.00026711 (cit. on p. 55).

[121] Patrice Hodonou Avogbe et al. "Urinary TERT promoter mutations as non-invasive biomarkers for the comprehensive detection of urothelial cancer". In: *EBioMedicine* 44 (June 2019), pp. 431–438. ISSN: 23523964. DOI: 10.1016/j.ebiom.2019.05.004 (cit. on p. 55).

[122] Katherine A. Hoadley et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin". In: *Cell* 158.4 (Aug. 2014), pp. 929–944. ISSN: 10974172. DOI: 10.1016/j.cell.2014.06.049 (cit. on pp. 55, 87).

[123] Muaiad Kittaneh, Alberto J. Montero, and Stefan Glück. "Molecular Profiling for Breast Cancer: A Comprehensive Review". In: *Biomarkers in Cancer* 5 (Jan. 2013), BIC.S9455. ISSN: 1179-299X. DOI: 10.4137/bic.s9455 (cit. on p. 56).

[124] Claudia Calabrese et al. "Genomic basis for RNA alterations in cancer". In: *Nature* 578.7793 (Feb. 2020), pp. 129–136. ISSN: 14764687. DOI: 10.1038/s41586-020-1970-0 (cit. on pp. 56, 142).

[125] Holly E. Barker and Clare L. Scott. "Preclinical rare cancer research to inform clinical trial design". In: *Nature Reviews Cancer* 19.9 (Sept. 2019), pp. 481–482. ISSN: 14741768. DOI: 10.1038/s41568-019-0172-2 (cit. on p. 56).

[126] *Rare Cancers Genomics*. URL: http://rarecancersgenomics.com/. Online. Accessed October 2020 (cit. on p. 56).

[127] Dorian R.A. Swarts et al. "Interobserver Variability for the WHO Classification of Pulmonary Carcinoids". In: *The American Journal of Surgical Pathology* 38.10 (Oct. 2014), pp. 1429–1436. ISSN: 0147-5185. DOI: 10.1097/PAS.0000000000000300 (cit. on p. 57).

[128] Charles M Rudin et al. "Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer". In: *Nature Genetics* 44.10 (Oct. 2012), pp. 1111–1116. ISSN: 1061-4036. DOI: 10.1038/ng.2405 (cit. on pp. 57, 90).

[129] Pierre Courtiol et al. "Deep learning-based classification of mesothelioma improves prediction of patient outcome". In: *Nature Medicine* 25.10 (Oct. 2019), pp. 1519–1525. ISSN: 1546170X. DOI: 10.1038/s41591-019-0583-3 (cit. on p. 62).

[130] Yu Fu et al. "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis". In: *bioRxiv* 44.0 (Feb. 2019), p. 813543. DOI: 10.1101/813543 (cit. on p. 62).

[131] Giuseppe Pelosi et al. "Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm". In: *Virchows Archiv* 472.4 (Apr. 2018), pp. 567–577. ISSN: 0945-6317. DOI: 10.1007/s00428-018-2307-3 (cit. on p. 63).

[132] Akash Mitra et al. "Spatially resolved analyses link genomic and immune diversity and reveal unfavorable neutrophil activation in melanoma". In: *Nature Communications* 11.1 (Dec. 2020), pp. 1–18. ISSN: 20411723. DOI: 10.1038/s41467-020-15538-9 (cit. on p. 63).

[133] Michael S. Lawrence et al. "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484 (Jan. 2014), pp. 495–501. ISSN: 00280836. DOI: 10.1038/nature12912 (cit. on p. 87).

[134] Katherine A. Hoadley et al. "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer". In: *Cell* 173.2 (Apr. 2018), pp. 291–304. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2018.03.022 (cit. on pp. 87, 88).

[135] Hamid Bolouri, Lue Ping Zhao, and Eric C. Holland. "Big data visualization identifies the multidimensional molecular landscape of human gliomas". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.19 (May 2016), pp. 5394–5399. ISSN: 10916490. DOI: 10.1073/pnas.1601591113 (cit. on p. 88).

[136] Yulia Newton et al. "TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal". In: *Cancer Research* 77.21 (Nov. 2017), e111–e114. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-17-0580 (cit. on pp. 88, 96, 97).

[137] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016), pp. 1–9. ISSN: 20524463. DOI: 10.1038/sdata.2016.18 (cit. on p. 88).

[138] Alessandra Fabbri et al. "Thymus neuroendocrine tumors with CTNNB1 gene mutations, disarrayed ß-catenin expression, and dual intra-tumor Ki-67 labeling index compartmentalization challenge the concept of secondary high-grade neuroendocrine tumor: a paradigm shift". In: *Virchows Archiv* 471.1 (July 2017), pp. 31–47. ISSN: 14322307. DOI: 10.1007/s00428-017-2130-2 (cit. on p. 89).

[139] N. Alcala et al. "Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids". In: *Nature Communications* 10.1 (Dec. 2019), p. 3407. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11276-9 (cit. on pp. 89, 90, 94, 97).

[140] Saurabh V. Laddha et al. "Integrative genomic characterization identifies molecular subtypes of lung carcinoids". In: *Cancer Research* 79.17 (2019), pp. 4339–4347. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-19-0214 (cit. on pp. 90, 94, 97).

[141] Evan W Floden Pablo Prieto Barja Emilio Palumbo Cedric Notredame Paolo Di Tommaso Maria Chatzou. "Nextflow enables reproducible computational workflows". In: *Nature Biotechnology* 35.4 (2017), pp. 316–319. DOI: 10.1038/nbt.3820 (cit. on pp. 90, 125).

[142] Paolo Di Tommaso et al. "The impact of Docker containers on the performance of genomic pipelines". In: *PeerJ* 2015.9 (2015), pp. 1–10. ISSN: 21678359. DOI: 10.7717/peerj.1273 (cit. on p. 90).

[143] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. "Singularity: Scientific containers for mobility of compute". In: *PLoS ONE* 12.5 (2017), pp. 1–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0177459 (cit. on p. 90).

[144] *IARC bioinformatics platform · GitHub*. URL: https://github.com/IARCbioinfo. Online. Accessed October 2020 (cit. on pp. 90, 95).

[145] Leland McInnes. *Basic UMAP Parameters — umap 0.4 documentation*. URL: https://umap-learn.readthedocs.io/en/latest/parameters.html. Online. Accessed October 2020 (cit. on p. 92).

[146] Rafael Messias Martins, Rosane Minghim, and AC Telea. "Explaining Neighborhood Preservation for Multidimensional Projections". In: *Computer Graphics & Visual Computing (CGVC)*. The Eurographics Association, 2015. ISBN: 978-3-905674-94-1. DOI: 10.2312/cgvc.20151234 (cit. on p. 93).

[147] *Nextjournal*. URL: https://nextjournal.com/explore. Online. Accessed October 2020 (cit. on p. 95).

[148] Andy Coenen and Adam Pearce. *Understanding UMAP*. URL: https://pair-code.github.io/understanding-umap/. Online. Accessed October 2020 (cit. on p. 98).

[149] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. "Density-Preserving Data Visualization Unveils Dynamic Patterns of Single-Cell Transcriptomic Variability". In: *bioRxiv* (May 2020), p. 2020.05.12.077776. DOI: 10.1101/2020.05.12.077776 (cit. on p. 98).

[150] Ben Langmead and Abhinav Nellore. "Cloud computing for genomic data analysis and collaboration". In: *Nature Reviews Genetics* 19.4 (Apr. 2018), pp. 208–219. ISSN: 14710064. DOI: 10.1038/nrg.2017.113 (cit. on pp. 98, 99, 152).

[151] Izumi Hinkson. *Genomics Cloud Pilots Expand Data Access - National Cancer Institute*. URL: https://www.cancer.gov/about-nci/organization/ccg/blog/2017/cloud-pilots-democratize-data. Online. Accessed October 2020 (cit. on p. 99).

[152] Brendan Maher. "Personal genomes: The case of the missing heritability". In: *Nature* 456.7218 (Nov. 2008), pp. 18–21. ISSN: 14764687. DOI: 10.1038/456018a (cit. on p. 111).

[153] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1 (July 2017), pp. 5–22. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2017.06.005 (cit. on p. 111).

[154] Patrick Turley et al. "Multi-trait analysis of genome-wide association summary statistics using MTAG". In: *Nature Genetics* 50.2 (Feb. 2018), pp. 229–237. ISSN: 1061-4036. DOI: 10.1038/s41588-017-0009-4 (cit. on p. 111).

[155] Brendan Bulik-Sullivan et al. "An atlas of genetic correlations across human diseases and traits". In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1236–1241. ISSN: 1061-4036. DOI: 10.1038/ng.3406 (cit. on p. 111).

[156] Minjie Chu et al. "A genome-wide gene-gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese". In: *Carcinogenesis* 35.3 (2014), pp. 572–577. ISSN: 14602180. DOI: 10.1093/carcin/bgt400 (cit. on p. 111).

[157] Amit V. Khera et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature Genetics* (Aug. 2018), p. 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0183-z (cit. on p. 111).

[158] Yan Dora Zhang et al. "Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers". In: *Nature Communications* 11.1 (Dec. 2020), p. 3353. ISSN: 2041-1723. DOI: 10.1038/s41467-020-16483-3 (cit. on pp. 111, 146).

[159] Hilary K. Finucane et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1228–1235. ISSN: 15461718. DOI: 10.1038/ng.3404 (cit. on p. 111).

[160] Stephen Burgess et al. "Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors". In: *European Journal of Epidemiology* 30.7 (July 2015), pp. 543–552. ISSN: 0393-2990. DOI: 10.1007/s10654-015-0011-z (cit. on p. 111).

[161] Michael Wainberg et al. "Opportunities and challenges for transcriptome-wide association studies". In: *Nature Genetics* 51.4 (Apr. 2019), pp. 592–599. ISSN: 15461718. DOI: 10.1038/s41588-019-0385-z (cit. on p. 112).

[162] The GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (Sept. 2020), pp. 1318–1330. ISSN: 0036-8075. DOI: 10.1126/science.aaz1776 (cit. on p. 112).

[163] Shuji Ogino et al. "Molecular pathological epidemiology of colorectal neoplasia: An emerging transdisciplinary and interdisciplinary field". In: *Gut* 60.3 (Mar. 2011), pp. 397–411. ISSN: 00175749. DOI: 10.1136/gut.2010.217182 (cit. on p. 112).

[164] Shuji Ogino, Charles S. Fuchs, and Edward Giovannucci. "How many molecular subtypes? Implications of the unique tumor principle in personalized medicine". In: *Expert Review of Molecular Diagnostics* 12.6 (July 2012), pp. 621–628. ISSN: 14737159. DOI: 10.1586/erm.12.46 (cit. on p. 112).

[165] Hannah Carter et al. "Interaction landscape of inherited polymorphisms with somatic events in cancer". In: *Cancer Discovery* 7.4 (Apr. 2017), pp. 410–423. ISSN: 21598290. DOI: 10.1158/2159-8290.CD-16-1045 (cit. on p. 112).

[166] Serena Nik-Zainal et al. "Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer". In: *Nature Genetics* 46.5 (2014), pp. 487–491. ISSN: 15461718. DOI: 10.1038/ng.2955 (cit. on p. 112).

[167] Candace D. Middlebrooks et al. "Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors". In: *Nature Genetics* 48.11 (Nov. 2016), pp. 1330–1338. ISSN: 15461718. DOI: 10.1038/ng.3670 (cit. on p. 112).

[168] Rayjean J. Hung et al. "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25". In: *Nature* 452.7187 (Apr. 2008), pp. 633–637. ISSN: 14764687. DOI: 10.1038/nature06885 (cit. on pp. 113, 135).

[169] Christopher I. Amos et al. "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1". In: *Nature Genetics* 40.5 (May 2008), pp. 616–622. ISSN: 10614036. DOI: 10.1038/ng.109 (cit. on pp. 113, 135).

[170] Marcus R. Munafò et al. "Association Between Genetic Variants on Chromosome 15q25 Locus and Objective Measures of Tobacco Exposure". In: *JNCI: Journal of the National Cancer Institute* 104.10 (May 2012), pp. 740–748. ISSN: 1460-2105. DOI: 10.1093/jnci/djs191 (cit. on p. 113).

[171] Thorgeir E. Thorgeirsson and Kari Stefansson. "Commentary: Gene-environment interactions and smoking-related cancers". In: *International Journal of Epidemiology* 39.2 (Apr. 2010), pp. 577–579. ISSN: 03005771. DOI: 10.1093/ije/dyp385 (cit. on p. 113).

[172] Yalei Zhang et al. "Chromosome 15q25 (CHRNA3-CHRNB4) Variation Indirectly Impacts Lung Cancer Risk in Chinese Males". In: *PLOS ONE* 11.3 (Mar. 2016). Ed. by Raymond Niaura, e0149946. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149946 (cit. on p. 113).

[173] Yufei Wang et al. "Common 5p15.33 and 6p21.33 variants influence lung cancer risk". In: *Nature Genetics* 40.12 (Dec. 2008), pp. 1407–1409. ISSN: 10614036. DOI: 10.1038/ng.273 (cit. on p. 113).

[174] James D. McKay et al. "Lung cancer susceptibility locus at 5p15.33". In: *Nature Genetics* 40.12 (2008), pp. 1404–1406. ISSN: 15461718. DOI: 10.1038/ng.254 (cit. on p. 113).

[175] Maria Teresa Landi et al. "A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma". In: *The American Journal of Human Genetics* 85 (2009), pp. 679–691. DOI: 10.1016/j.ajhg.2009.09.012 (cit. on pp. 113, 135).

[176] Stephen S. Hecht. "Tobacco carcinogens, their biomarkers and tobacco-induced cancer". In: *Nature Reviews Cancer* 3.10 (2003), pp. 733–744. ISSN: 1474175X. DOI: 10.1038/nrc1190 (cit. on p. 113).

[177] Ludmil B Alexandrov et al. "Mutation signatures associated with tobacco smoking in human cancer". In: *Science* 354.6312 (2016), pp. 618–622. ISSN: 1095-9203. DOI: 10.1126/science.aag0299 (cit. on pp. 113, 137).

[178] Jonathan Marchini and Bryan Howie. "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (June 2010), pp. 499–511. ISSN: 14710056. DOI: 10.1038/nrg2796 (cit. on p. 114).

[179] NCI Genomic Data Commons. *GDC Data Transfer Tool*. URL: https://gdc.cancer.gov/access-data/gdc-data-transfer-tool. Online. Accessed October 2020 (cit. on p. 114).

[180] Mariza de Andrade et al. "Evaluating the Influence of Quality Control Decisions and Software Algorithms on SNP Calling for the Affymetrix 6.0 SNP Array Platform". In: *Human Heredity* 71.4 (Sept. 2011), pp. 221–233. ISSN: 0001-5652. DOI: 10.1159/000328843 (cit. on p. 115).

[181] *Affymetrix Support by Product for Genome-Wide Human SNP Array 6.0*. URL: http://www.affymetrix.com/support/technical/byproduct.affx?product=genomewidesnp_6. Online. Accessed October 2020 (cit. on p. 115).

[182] Jianfang Liu et al. "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics." In: *Cell* 173.2 (Apr. 2018), pp. 400–416. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.02.052 (cit. on p. 115).

[183] *McCarthy Tools*. URL: https://www.well.ox.ac.uk/~wrayner/tools/#Checking. Online. Accessed October 2020 (cit. on pp. 115, 117).

[184] David H. Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* (2009). ISSN: 10889051. DOI: 10.1101/gr.094052.109 (cit. on p. 115).

[185] *PLINK. The Phase 2 HapMap as a PLINK fileset*. URL: http://zzz.bwh.harvard.edu/plink/res.shtml. Online. Accessed October 2020 (cit. on p. 115).

[186] Kai Yu et al. "Population substructure and control selection in genome-wide association studies". In: *PLoS ONE* 3.7 (July 2008). ISSN: 19326203. DOI: 10.1371/journal.pone.0002551 (cit. on pp. 115, 121).

[187] Po Ru Loh, Pier Francesco Palamara, and Alkes L. Price. "Fast and accurate long-range phasing in a UK Biobank cohort". In: *Nature Genetics* 48.7 (July 2016), pp. 811–816. ISSN: 15461718. DOI: 10.1038/ng.3571 (cit. on p. 120).

[188] Susan Fairley et al. "The International Genome Sample Resource (IGSR) collection of open human genomic variation resources". In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D941–D947. ISSN: 13624962. DOI: 10.1093/nar/gkz836 (cit. on p. 120).

[189] Sayantan Das et al. "Next-generation genotype imputation service and methods". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1284–1287. ISSN: 15461718. DOI: 10.1038/ng.3656 (cit. on p. 120).

[190] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nat. Genet* 38.8 (2006), pp. 904–909. ISSN: 1061-4036. DOI: 10.1038/ng1847 (cit. on p. 121).

[191] *Michigan Imputation Server*. URL: https://imputationserver.sph.umich.edu/index.html#!. Online. Accessed October 2020 (cit. on pp. 121, 152).

[192] *TOPMed Imputation Server*. URL: https://imputation.biodatacatalyst.nhlbi.nih.gov/#!. Online. Accessed October 2020 (cit. on pp. 121, 152).

[193] Buhm Han and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies". In: *American Journal of Human Genetics* 88.5 (May 2011), pp. 586–598. ISSN: 00029297. DOI: 10.1016/j.ajhg.2011.04.014 (cit. on p. 122).

[194] Nasim Mavaddat et al. "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes". In: *The American Journal of Human Genetics* 104.1 (Jan. 2019), pp. 21–34. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2018.11.002 (cit. on p. 123).

[195] Matthew Warren. "The approach to predictive medicine that is taking genomics research by storm". In: *Nature* 562.7726 (Oct. 2018), pp. 181–183. ISSN: 0028-0836. DOI: 10.1038/d41586-018-06956-3 (cit. on pp. 123, 150).

[196] Mengzhen Liu et al. "Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use". In: *Nature Genetics* (Jan. 2019), p. 1. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0307-5 (cit. on pp. 123, 126).

[197] Xia Jiang et al. "Shared heritability and functional enrichment across six solid cancers". In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 20411723. DOI: 10.1038/s41467-018-08054-4 (cit. on p. 123).

[198] Corina Lesseur et al. "Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer". In: *Nature Genetics* 48.12 (Dec. 2016), pp. 1544–1550. ISSN: 15461718. DOI: 10.1038/ng.3685 (cit. on p. 123).

[199] Shing Wan Choi, Timothy Shin Heng Mak, and Paul F. O'Reilly. "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature Protocols* (July 2020), pp. 1–14. ISSN: 17502799. DOI: 10.1038/s41596-020-0353-1 (cit. on p. 123).

[200] Kyle Ellrott et al. "Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines." In: *Cell systems* 6.3 (Mar. 2018), pp. 271–281. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.03.002 (cit. on p. 124).

[201] Felix Dietlein et al. "Identification of cancer driver genes based on nucleotide context". In: *Nature Genetics* 52.2 (Feb. 2020), pp. 208–218. ISSN: 15461718. DOI: 10.1038/s41588-019-0572-y (cit. on p. 125).

[202] Francis Blokzijl et al. "MutationalPatterns: comprehensive genome-wide analysis of mutational processes". In: *Genome Medicine* 10.1 (Dec. 2018), p. 33. ISSN: 1756-994X. DOI: 10.1186/s13073-018-0539-0 (cit. on p. 125).

[203] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. "GSVA: gene set variation analysis for microarray and RNA-Seq data". In: *BMC Bioinformatics* 14.1 (Jan. 2013), p. 7. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-7 (cit. on p. 126).

[204] Jeff Kiefer et al. "Abstract 3589: A systematic approach toward gene annotation of the hallmarks of cancer". In: *Cancer Research*. Vol. 77. 13 Supplement. American Association for Cancer Research (AACR), July 2017, pp. 3589–3589. DOI: 10.1158/1538-7445.am2017-3589 (cit. on p. 126).

[205] Matthew E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (Apr. 2015), e47–e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007 (cit. on p. 126).

[206] Debbie A. Lawlor et al. "Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology". In: *Statistics in Medicine* 27.8 (Apr. 2008), pp. 1133–1163. ISSN: 02776715. DOI: 10.1002/sim.3034 (cit. on p. 126).

[207] D A Bennett and M V Holmes. "Mendelian randomisation in cardiovascular research: an introduction for clinicians". In: *Heart* 0 (2017), pp. 1–8. DOI: 10.1136/heartjnl-2016-310605 (cit. on p. 126).

[208] Neil M Davies, Michael V Holmes, and George Davey Smith. "Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians." In: *BMJ (Clinical research ed.)* 362 (July 2018), k601. ISSN: 1756-1833. DOI: 10.1136/bmj.k601 (cit. on pp. 126, 127).

[209] Gibran Hemani et al. "The MR-Base platform supports systematic causal inference across the human phenome". In: *eLife* 7 (May 2018). ISSN: 2050-084X. DOI: 10.7554/eLife.34408 (cit. on p. 127).

[210] Xuanyao Liu, Yang I. Li, and Jonathan K. Pritchard. "Trans Effects on Gene Expression Can Drive Omnigenic Inheritance". In: *Cell* 177.4 (2019), pp. 1022–1034. ISSN: 10974172. DOI: 10.1016/j.cell.2019.04.014 (cit. on p. 131).

[211] *CHRNA5 - HumanBase*. URL: https://hb.flatironinstitute.org/gene/1138 (cit. on pp. 134, 136).

[212] Joshua D Campbell et al. "Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas". In: *Nature Genetics* 48.6 (June 2016), pp. 607–616. ISSN: 1061-4036. DOI: 10.1038/ng.3564 (cit. on p. 135).

[213] H Lee, JK Aronson, and D Nunan. *Catalogue of bias collaboration*. 2019 (cit. on p. 136).

[214] Yohan Bossé et al. "Transcriptome-wide association study reveals candidate causal genes for lung cancer". In: *International Journal of Cancer* 146.7 (Apr. 2020), pp. 1862–1878. ISSN: 10970215. DOI: 10.1002/ijc.32771 (cit. on p. 136).

[215] L. Duncan et al. "Analysis of polygenic risk score usage and performance in diverse human populations". In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 20411723. DOI: 10.1038/s41467-019-11112-0 (cit. on p. 137).

[216] *Mutographs project*. URL: https://www.mutographs.org/mutographs-project/. Online. Accessed October 2020 (cit. on p. 138).

[217] Mariam Jamal-Hanjani et al. "Tracking the Evolution of Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 376.22 (June 2017), pp. 2109–2121. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1616288 (cit. on p. 138).

[218] Loïc Le Marchand et al. "Smokers with the CHRNA lung cancer-associated variants are exposed to higher levels of nicotine equivalents and a carcinogenic tobacco-specific nitrosamine". In: *Cancer Research* 68.22 (Nov. 2008), pp. 9137–9140. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-08-2271 (cit. on p. 138).

[219] Robyn E. Wootton et al. "Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study". In: *Psychological Medicine* 50.14 (2020), 2435–2443. DOI: 10.1017/S0033291719002678 (cit. on p. 138).

[220] Tiago C. Silva et al. "ELmer v.2: An r/bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles". In: *Bioinformatics* 35.11 (2019), pp. 1974–1977. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty902 (cit. on p. 142).

[221] Nature. "The era of massive cancer sequencing projects has reached a turning point". In: *Nature* 578.7793 (Feb. 2020), pp. 7–8. ISSN: 14764687. DOI: 10.1038/d41586-020-00308-w (cit. on p. 143).

[222] Sara R. Rashkin et al. "Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts". In: *Nature Communications* 11.1 (Dec. 2020), pp. 1–14. ISSN: 20411723. DOI: 10.1038/s41467-020-18246-6 (cit. on p. 143).

[223] Robert M. Maier et al. "Improving genetic prediction by leveraging genetic correlations among human diseases and traits". In: *Nature Communications* 9.1 (Dec. 2018), p. 989. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02769-6 (cit. on p. 143).

[224] Alexander S. Rich and Todd M. Gureckis. "Lessons for artificial intelligence from the study of natural stupidity". In: *Nature Machine Intelligence* 1.4 (Apr. 2019), pp. 174–180. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0038-z (cit. on pp. 144, 151).

[225] Mohammad Peikari et al. "A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification." In: *Scientific reports* 8.1 (May 2018), p. 7193. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24876-0 (cit. on p. 144).

[226] Boyu Lyu and Anamul Haque. "Deep Learning Based Tumor Type Classification Using Gene Expression Data". In: *bioRxiv* (July 2018), p. 364323. DOI: 10.1101/364323 (cit. on p. 145).

[227] Joseph M. De Guia, Madhavi Devaraj, and Carson K. Leung. "DeepGX: Deep learning using gene expression for cancer classification". In: *Association for Computing Machinery*. Aug. 2019, pp. 913–920. ISBN: 9781450368681. DOI: 10.1145/3341161.3343516 (cit. on p. 145).

[228] Hannah L. Nicholls et al. "Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci". In: *Frontiers in Genetics* 11 (Apr. 2020). ISSN: 16648021. DOI: 10.3389/fgene.2020.00350 (cit. on p. 145).

[229] Christina B Azodi, Jiliang Tang, and Shin-Han Shiu. "Opening the Black Box: Interpretable Machine Learning for Geneticists". In: *Trends in Genetics* 36 (2020), pp. 442–455. DOI: 10.1016/j.tig.2020.03.005 (cit. on p. 145).

[230] Linda Kachuri et al. "Integration of polygenic risk scores with modifiable risk factors improves risk prediction: results from a pan-cancer analysis". In: *bioRxiv* (Sept. 2020), p. 2020.01.28.922088. DOI: 10.1101/2020.01.28.922088 (cit. on p. 146).

[231] Charles Swanton. "Take lessons from cancer evolution to the clinic". In: *Nature* 581.7809 (May 2020), pp. 382–383. ISSN: 14764687. DOI: 10.1038/d41586-020-01347-z (cit. on p. 146).

[232] *ICGC ARGO - Introduction and Goals*. URL: https://www.icgc-argo.org/page/72/introduction-and-goals-. Online. Accessed October 2020 (cit. on p. 148).

[233] Gary Saunders et al. "Leveraging European infrastructures to access 1 million human genomes by 2022". In: *Nature Reviews Genetics* 20.11 (Nov. 2019), pp. 693–701. ISSN: 14710064. DOI: 10.1038/s41576-019-0156-9 (cit. on pp. 148, 153).

[234] Gouvernement.fr. *Plan "France médecine génomique 2025" : lancement des 2 premières plateformes*. URL: https://www.gouvernement.fr/partage/9344-plan-france-medecine-genomique-2025-lancement-des-2-premieres-plateformes. Online. Accessed October 2020 (cit. on p. 148).

[235] C. Turnbull. "Introducing whole-genome sequencing into routine cancer care: The Genomics England 100 000 Genomes Project". In: *Annals of Oncology* 29.4 (Apr. 2018), pp. 784–787. ISSN: 15698041. DOI: 10.1093/annonc/mdy054 (cit. on p. 148).

[236] Genomics England. *GeCIP Domains*. URL: https://www.genomicsengland.co.uk/about-gecip/gecip-domains/. Online. Accessed October 2020 (cit. on p. 149).

[237] David Tamborero et al. "Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations". In: *Genome Medicine* 10.1 (Mar. 2018), p. 25. ISSN: 1756994X. DOI: 10.1186/s13073-018-0531-8 (cit. on p. 149).

[238] Jian Carrot-Zhang et al. "Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer". In: *Cancer Cell* 37.5 (2020), pp. 639–654. ISSN: 18783686. DOI: 10.1016/j.ccell.2020.04.012 (cit. on p. 151).

[239] Amy L. McGuire et al. "The road ahead in genetics and genomics". In: *Nature Reviews Genetics* (Aug. 2020), pp. 1–16. ISSN: 1471-0056. DOI: 10.1038/s41576-020-0272-6 (cit. on p. 151).

[240] Philip A. Ewels et al. "The nf-core framework for community-curated bioinformatics pipelines". In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 276–278. ISSN: 15461696. DOI: 10.1038/s41587-020-0439-x (cit. on p. 152).

[241] Cancer Genome Collaboratory. *Cloud Computing for Big Data Genomics*. URL: https://cancercollaboratory.org/. Online. Accessed October 2020 (cit. on p. 152).

[242] Xiaoqing Guan et al. "CVCDAP: an integrated platform for molecular and clinical analysis of cancer virtual cohorts". In: *Nucleic acids research* 48.W1 (July 2020), W463–W471. ISSN: 13624962. DOI: 10.1093/nar/gkaa423 (cit. on p. 152).

[243] Mark Phillips et al. "Genomics: data sharing needs an international code of conduct". In: *Nature* 578.7793 (Feb. 2020), pp. 31–33. ISSN: 14764687. DOI: 10.1038/d41586-020-00082-9 (cit. on p. 153).

[244] *Beyond One Million Genomes (B1MG) project*. URL: https://b1mg-project.eu/. Online. Accessed October 2020 (cit. on p. 153).

[245] *Global Alliance for Genomics and Health (GA4GH)*. URL: https://www.ga4gh.org/aboutus/. Online. Accessed October 2020 (cit. on p. 153).

Acronyms

AF Alternative frequency. 118–120

AgeInit age of initiation. 126, 131

APT Affymetrix Power Tools. 115

ATAC-seq Assay of Transposase Accessible Chromatin sequencing. 33

B1MG Beyond 1 Million Genomes. 153

BAM Binary Alignment Map. 91

BCF Binary Variant Call Format,. 120

BMI Body Mass Index. 137

BRCA Breast Invasive Carcinoma. 115

cDNA complementary DNA. 30

CGC Cancer Genomics Cloud. 37, 99

CGI Cancer Genome Interpreter. 149

ChiP-Seq Chromatin immunoprecipitation Sequencing. 33

CHRNA3 Cholinergic Receptor Nicotinic Alpha 3 Subunit. 113

CHRNA5 Cholinergic Receptor Nicotinic Alpha 5 Subunit. 113, 127, 134, 136

CHRNB4 Cholinergic Receptor Nicotinic Beta 4 Subunit. 113

CML chronic myelogenous leukemia. 22

COPD chronic obstructive pulmonary disease. 136, 142

COSMIC Catalogue Of Somatic Mutations In Cancer. 40, 125

CPD cigarettes per day. 126, 131, 138

CpG cytosine–phosphate–guanine. 20, 30, 33

ctDNA circulating tumor DNA. 41, 55

CVCDAP Cancer Virtual Cohort Discovery Analysis Platform. 152

dbGAP Database of Genotypes And Phenotypes. 36

DIABLO Data Integration Analysis for Biomarker discovery using Latent cOmponents. 52, 137

DNA Deoxyribonucleic acid. 18–28, 31, 33, 40, 41, 90, 97, 113, 114, 127, 136, 137, 142, 143

DR dimensionality reduction. 48, 50–52

EDA exploratory data analyses. 43

EGA European-Genome Phenome Archive. 36, 90

EGFR Epidermal Growth Factor Receptor. 40, 55

ELMER Enhancer Linking by Methylation/Expression Relationships. 142

eQTL expression quantitative trait loci. 112, 135

FAIR Findable, Accessible, Interoperable, Reusable. 88

FEV Forced Expiratory Volume. 143

FFPE Formalin-Fixed Paraffin-Embedded. 97, 98

GA4GH Global Alliance for Genomics and Health. 153

GATK Genome Analysis Toolkit. 91

GDC Genomic Data Common. 37, 99, 114

GDPR General Data Protection Regulation. 153

GeCIP Genomics England Clinical Interpretation Partnership. 149

GEO Gene Expression Omnibus. 36, 90

GSCAN GWAS Sequencing Consortium of Alcohol and Nicotine. 123, 126, 131

GSEA Gene Set Enrichment Analyses. 59, 61

GSVA Gene Set Variation Analysis. 125, 126, 136

GTEx Genotype-Tissue Expression. 112

gw-SNPs genome-wide significant SNPs. 123, 124, 127

GWAS Genome-Wide Association Studies. 28, 29, 36, 39, 53, 111–114, 122, 123, 127, 128, 135, 137, 143, 145, 146, 151

GWAx family history GWAS. 122

HapMap Haplotype Map project. 29, 115

HGP Human Genome Project. 26

IARC International Agency for Research on Cancer. 17

ICA Independent Component Analysis. 49

ICGC International Cancer Genome Consortium. 35–37, 56, 87, 137, 148

ICGC-ARGO Accelerate Research in Genomic Oncology. 148

IGSR International Genome Sample Resource. 120

indels insertions or deletions. 22, 26

IREB2 Iron Responsive Element Binding Protein 2. 136

ISB-CGC ISB Cancer Genomics Cloud. 37

IVW Inverse Variance Weighted. 127, 131

LCNEC Large Cell Neuroendocrine Carcinoma. 39, 41, 56–63, 88, 90, 94, 96, 143

LD linkage disequilibrium. 29

LNEN Lung Neuroendocrine Neoplasm. 53, 57–60, 62, 88, 89, 94, 97, 98, 141, 143, 144, 147–149, 153

LOH loss of heterozygosity. 23, 28

LUAD Lung Adenocarcinomas. 38–40, 55, 112, 113, 124, 125, 129, 130, 133, 135, 147, 148

lung NENs lung neuroendocrine neoplasms. 53, 56, 57, 60

LUSC Lung Squamous Cell Carcinomas. 38–40, 112, 113, 124, 125, 129, 130, 135, 147, 148

MAF Minor Allele Frequency. 117, 120

maf Mutation Annotation Format. 124, 125

MC3 Multi-Center Mutation Calling in Multiple Cancers. 124

MHC Major Histocompatibility Complex. 61

MI Moran index. 93, 95

miRNAs micro RNAs. 20, 34

ML machine learning. 43, 45, 58–62

MOFA Multi-Omics Factor Analysis. 51, 58, 59, 61, 62, 142

MR Mendelian Randomization. 122, 126, 127, 131–133, 136–138

MRI Magnetic Resonance Imaging. 36

mRNAs messenger RNAs. 20, 27, 30, 33

MSI Microsatellite Instability. 137

NCI National Cancer Institute. 37

NEN Neuroendocrine neoplasm. 63, 89, 90

NGS Next Generation Sequencing. 31, 32, 34, 40

NIH National Institutes of Health. 34, 37

NMF Non-negative Matrix factorization. 49

NNLS non-negative least squares. 125

NSCLC Non Small Cell Lung Cancer. 38, 40, 55, 141, 146

PC principal components. 93

PCA Principal Component Analysis. 43, 49, 51, 52, 59, 89, 91, 93, 95, 123

PCAWG Pan-Cancer Analysis of Whole Genomes. 35, 37, 137, 152

PCR Polymerase Chain Reaction. 32

PLS Partial Least Squares. 52, 123

PLS-DA PLS discriminant analysis. 52, 137

pre-mRNAs precursor messenger RNA. 20

PRS Polygenic Risk Scores. 111, 122–124, 126–131, 133, 135, 137, 146, 150, 151

QC quality controls. 90, 115

RCT Randomized Control Trials. 126

RNA Ribonucleic acid. 19–21, 32–34, 57, 90, 97, 142

RNA-Seq RNA Sequencing. 33, 48, 91, 94, 96, 125

RPPA Reverse-Phase Protein Array. 34

SCLC Small Cell Lung Cancer. 38–41, 55–58, 90, 94, 96

SD sequence difference. 93, 95

SmkCes smoking cessation. 126, 131

SmkInit smoking initiation. 126, 131

SNP Single Nucleotide Polymorphism. 22, 27–29, 31, 111, 112, 115, 117–124, 126–129, 131, 133, 135, 137, 138, 141

SNVs Single Nucleotide Variations. 22

SVM Support Vector Machines. 45

t-SNE t-Distributed Stochastic Neighbor Embedding. 50, 51, 98

TCGA The Cancer Genome Atlas. 34–37, 40, 56, 87, 88, 99, 112, 114–121, 123–125, 127, 136, 137, 148, 152

TRACERx Tracking Cancer Evolution through Therapy. 146

TRICL-ILCCO Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium. 122, 127