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

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

Affiliations

[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 AZ Maastricht, The Netherlands
[5] International Agency for Research on Cancer (IARC/WHO), Section of Epigenetics, 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 Platform, Cancer Research Centre of Lyon (CRCL), 28 Rue Laennec, 69008 Lyon, France
[9] Institute for Advanced Biosciences, Site Santé, Allée des Alpes, 38700 La Tronche, Grenoble, France
[10] Institute of Pathology and Molecular Pathology, University of Zurich, Schmelzbergstrasse 12 8091 Zurich, Switzerland
[11] Jena University Hospital, Bachstraße 18, 07743 Jena, Germany
[12] Bad Berka Institute of Pathology, Robert-Koch-Allee 9, 99438 Bad Berka, Germany
[13] Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, 6 West Derby Street, L7 8TX Liverpool, UK
[14] Pathology Institute, Hospices Civils de Lyon, University Claude Bernard Lyon 1, 59 Boulevard Pinel, 69677 BRON Cedex, France
[15] Caen University Hospital, CHU Caen, 3 avenue du Général Harris, 14076 Caen Cedex 5, France
[16] Department of Pathology, Centre Léon Bérard, 28, rue Laennec, 69373 Lyon Cedex 8, France
[17] St Vincent's Hospital and University of Melbourne, Victoria Parade, Fitzroy VIC 3065, Melbourne, Australia
[18] Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian 1, 20133 Milan, Italy
[19] Nancy Regional University Hospital, CHRU, CRB, INSERM U1256, 29 Avenue du Maréchal de Lattre de Tassigny, 54035 Nancy Cedex, France
[20] 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

[21] Drammen Hospital, Vestre Viken Health Trust, Vestre Viken HF, Postboks 800, 3004 Drammen, Norway
[22] Oslo University Hospital, Sognsvannsveien 20, 0372 Oslo, Norway
[23] Marie Lannelongue Hospital, 133 avenue de la Resistance, 92350 Le Plessis Robinson, France
[24] Fondazione IRCCS Casa Sollievo della Sofferenza, Viale Cappuccini 2, 71013 San Giovanni Rotondo FG, Italy
[25] Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria
[26] Department of Thoracopulmonary Pathology, Service of Pathology, Clinical Center of Serbia, Pasterova 2, Belgrade 11000, Serbia
[27] Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Joseph-Stelzmann-Straße 26, 50931 Cologne, Germany
[28] Department of Oncology and Hemato-Oncology, University of Milan, and Inter-Hospital Pathology Division, IRCCS Multimedica, Via Gaudenzio Fantoli, 16/15, 20138 Milan, Italy
[29] Department of Oncology, University of Turin, Via Santena 5, 10126 Torino, Italy
[30] Department of Immunity, Virus, and Inflammation, Cancer Research Centre of Lyon (CRCL), 28 Rue Laennec, 69008 Lyon, France
[31] Cancer Research Centre of Lyon (CRCL), Inserm U 1052, CNRS UMR 5286, Centre Léon Bérard, Université de Lyon, 28 Rue Laennec, 69008 Lyon, France
[32] University Lyon 1, Lyon, France; INSERM U932, Paris, France; Institut Curie, 26 Rue d'Ulm, 75005 Paris, France
[33] European Reference Network (ENR-EURACAN)

*These authors contributed equally
#These authors jointly supervised this work

Correspondence should be addressed to: fernandezcuestal@iarc.fr

Running title: Integrative analyses of lung neuroendocrine neoplasms

Abstract

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 with molecular and clinical features of the deadly LCNEC, validating the previously proposed molecular link between the low- and high-grade lung neuroendocrine neoplasms.

Introduction

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, which annual incidence 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% metastasize 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. Due 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 and 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 of 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 Figure 1**;

Supplementary Table 1). 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[4,11-15], for which pathological review had already been done.

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)[16] (MOFA LNEN; **Figure 1A; Supplementary Figures 2** and **3**;). We identified five latent factors explaining more than 2% of the variance in at least one dataset, 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 Figure 4**). Using consensus clustering on these two latent factors (which explained most of the variation and thus carried the most biological signal; **Supplementary Figures 5-7**; **Supplementary Tables 2-3**), we identified three clusters, each of them enriched for samples of one of the three histopathological types (**Figure 1A**). Cluster Carcinoid A was enriched for typical carcinoids (75%; Fisher's exact test p-value<$2.2x10^{-16}$); cluster Carcinoid B was enriched for atypical carcinoids (54%; Fisher's exact test p-value<$2.2x10^{-16}$) and male patients (79%; Fisher's exact test p-value=$1.6x10^{-9}$); and cluster LCNEC included 92% of the histopathological LCNEC (Fisher's exact test p-value<$2.2x10^{-16}$). Note that clustering based on LF1 to LF5, weighted by their proportion of variance explained, leads to the exact same clusters (**Supplementary Figure 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 datasets was available, thus maximizing the power of subsequent analyses (**Figure 1B**; **Supplementary Figure 9** for an alternative analysis based on both 'omic datasets 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 normalization, learned from the training set and applied to the test sample. To identify intermediate profiles, we defined a prediction category, Unclassified, for samples which probability ratio between the two most probable classes was close to 1. We present in **Figure 1B** the results for a cut-off ratio of 1.5, and show in **Supplementary Figure 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 (**Figure 1A**). Similarly, the majority of ML-predicted atypical carcinoids (87%) belonged to cluster Carcinoid B.

We selected the ML-prediction groups with more than 10 samples (gathering the unclassified samples in one single group) and compared their overall survival using Cox's proportional hazard model (coloured groups in **Figure 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 **Figure 1B-C**) with a 10-year overall survival similar to that of samples confirmed by ML as typical carcinoids (in black in **Figure 1B-C**; 88% and 89%, respectively; Wald test *p*-value=0.650); and the bad-prognosis group (atypical predicted as atypical, in red in **Figure 1B-C**) with a 10-year overall survival similar to that of samples confirmed by ML as LCNEC (in blue in **Figure 1B-C**; 27% and 19% respectively; Wald test *p*-value=0.574; see also **Supplementary Figure 11**). Machine-learning analyses based on other features such as combined expression and methylation data (**Supplementary Figure 9**), MOFA latent factors (**Supplementary Figure 12A**), and Principal Component Analyses (PCA) Principal Components explaining more than 2% of the variance (**Supplementary Figure 12B**), led to qualitatively similar results.

Atypical carcinoids with LCNEC molecular characteristics

Six atypical carcinoids clustered with LCNEC in the MOFA LNEN (supra-carcinoids; **Figure 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%, respectively; Wald test *p*-value=0.574; **Figure 2A**). The observed molecular link appear to be between supra-carcinoids and LCNEC rather than with SCLC, as shown by PCA and MOFA including expression data for 51 SCLC (**Supplementary Figure 6** and **Supplementary Figure 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 etiological 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)[17] (**Table 1**), and the tumour harboured a splicing *BAP1* somatic mutation (a gene frequently altered in mesothelioma)[18]. This sample showed the highest mutational load (37 damaging somatic mutations**; Supplementary Table 4**). Gene Set Enrichment Analyses (GSEA) of mutations in the hallmarks of cancer gene sets [19,20], showed a significant enrichment for the hallmark evading growth suppressor (*q*-value=0.0213; **Figure 2B**; **Supplementary Table 5**), while genome instability and mutation were significant at the 10% False-Discovery-Rate (FDR) threshold (*q*-value=0.0970; **Figure 2B**; **Supplementary Table 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 (**Figure 2C**; **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 (**Figure 2D**). Supra-carcinoids displayed higher levels of immune checkpoint genes (both receptors and ligands; **Figure 2E**), and also harboured generally higher expression levels of MHC class I and II genes (**Figure 2E**; **Supplementary Figure 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 Figure 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 (**Figure 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; **Figure 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 Figure 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 Table 6**). By contrast, no such association was observed in the MOFA performed only on carcinoids and SCLC samples (**Supplementary Figure 6C**; **Supplementary Figure 13C**; **Supplementary Table 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 (**Figure 3A**; **Supplementary Figure 16**; **Supplementary Table 4**). Apart from the above-mentioned genes, *ATM, PSIP1,* and *ROBO1* also showed some evidence, among others, for recurrent mutations in pulmonary carcinoids (**Figure 3A**). In addition to point mutations and small indels, the *ARID2*, *DOT1L*, and *ROBO1* genes were also altered by chimeric transcripts (**Figure 3B**). *MEN1* was also inactivated by genomic rearrangement in a carcinoid sample with a chromothripsis pattern affecting chromosomes 11 and 20 (**Figure 3C**). The full lists of somatically altered genes, chimeric transcripts, and genomic rearrangements are presented in **Supplementary Tables 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 Figure 4**). The molecular variation explained by LF3 appeared to capture different molecular profiles within cluster Carcinoid A (**Supplementary Figure 13B**).

We therefore undertook an additional MOFA restricted to pulmonary carcinoid samples only (MOFA LNET; **Figure 4A; Supplementary Figure 17**). This MOFA identified five latent factors that explained at least 2% of the variance in one dataset. 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 greater than 0.96; **Supplementary Figure 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 Figure 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 Figure 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 Table 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 leukocyte 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)[21] (**Supplementary Table 6**), and all CXC chemokines (such as IL8, CXCL1, CXCL3, and CXCL5, known to attract neutrophils)[22], 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 Figure 19**): cluster Carcinoid A (particularly A1) was significantly enriched in dendritic cells, and cluster Carcinoid B, in monocytes (**Figure 4B**, **upper panel**). As monocytes can differentiate into dendritic cells in a favourable environment[23], we assessed the levels of *LAMP3* and *CD1A* dendritic-cells markers[24], and found that samples in cluster Carcinoid A1 presented high expression levels of these genes (**Figure 4B**, **lower panel**), implying that this cluster was indeed enriched for dendritic cells. We pursued this further by assessing 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 (**Figure 4C**; **Supplementary Table 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 Table 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; **Figure 4D**), which contrasts with the low levels in other carcinoids (median FPKM=0 for both *UGT2A3* and *UGT2B*; **Figure 4D**), LCNEC (median FPKM=0 and 1.2 for *UGT2A3* and *UGT2B*; **Supplementary Figure 20**) and SCLC (median FPKM=0 and 0.3 for *UGT2A3* and *UGT2B*; **Supplementary Figure 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, genes which expression levels defined a given group of samples), corresponding promoter methylation profiles (**Figure 5A**; **Supplementary Table 10**), and their somatic mutational patterns (**Figure 3A; Figure 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 Figure 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 (**Figure 5A**). One of the top correlations was found for *HNF1A* and *HNF4A* homeobox genes (**Supplementary Figure 22**), which were strongly down-regulated in cluster Carcinoid A1 samples (**Supplementary Figure 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 Table 11**). These two genes have been reported as having a role in the transcriptional regulation of *ANGPTL3*, CYP, and UGT genes[25], 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*) (**Figure 5A**; **Supplementary Table 10**), similar to SCLC and LCNEC (**Figure 5B**); however, the expression levels of NOTCH genes did not differ between the different groups (**Supplementary Figure 24**). The supra-carcinoids were negative for *DLL3* expression (**Figure 5B**), and had generally high expression levels of *NOTCH1-3* (**Supplementary Figure 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 (**Figure 4D**; **Supplementary Table 9**), while in the other 12 samples DLL3 was strikingly absent (**Figure 4D**; **Supplementary Table 9**). Furthermore, we found a correlation between the protein levels of DLL3 and CD1A (Pearson test *p*-value=0.00034; **Supplementary Figure 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) (**Figure 5A-B**;

Supplementary Table 10). This cluster also contained the four samples with somatic mutations in the eukaryotic translation initiation factor 1A X-linked (*EIF1AX*) gene (**Figure 4A**). Concordantly, samples with *EIF1AX* mutations had significantly higher coordinates on the MOFA LNET LF2 (t-test *p*-value=0.0342).

As expected based on **Figure 4D**, several UGT genes were core-DEGs of cluster Carcinoid B (**Figure 5A**). Also, accordingly with the worse survival of patients in this cluster (**Figure 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) (**Figure 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) (**Figure 5B**). Interestingly, the SCLC-combined LCNEC sample (S00602) that clustered with the pulmonary carcinoids in the MOFA LNEN (**Figure 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 Table 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=7x10^{-6}; **Figure 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 Figure 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 [19,20]; 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 (**Figure 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 (**Figure 5D**; **Supplementary Table 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 (**Figure 1C**, **upper panel**; **Supplementary Figure 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 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 Figure 11B-C**).

An overview of the different molecular groups of pulmonary carcinoids and their most relevant characteristics is displayed in **Figure 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 histopathological 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 (**Figure 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[26]. 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[27]. Overexpression of *ASCL1* and *DLL3* is a characteristic of the SCLC of the classic subtype[27] and the 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 humanized monoclonal antibody in high–DLL3-expressing SCLCs and LCNECs[28], and additional clinical trials are ongoing in other cancer types.

The second group (cluster A2) harboured recurrent somatic mutations in *EIF1AX*, and showed down-regulation 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[29], 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[30], 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[31]. 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[32]. In cancer, this pathway mainly suppresses tumour progression by regulating invasion, migration, and apoptosis, and therefore, is often down-regulated in many cancer types[29]. 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[29]. Potential clinical avenues to this finding exist, especially the on-going 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[33]. Cluster B was also characterised by recurrent somatic mutations in *MEN1*, the most frequently altered gene in pulmonary carcinoids and pancreatic NETs[34], 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[35], 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[15,36], 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 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 *versus* 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 PDL1 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 up-regulation 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 characterize 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-characterized 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

Clinical data

Collected clinical data included age (in years), sex (male or female), smoking status (never smoker, former smoker, passive smoker, and current smoker), UICC/AJCC 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, Nat comm (2013), George et al, Nat Comm (2017), and George et al, Nature, (2015)[11,12,14]. In order to improve the power of the statistical analyses, we regrouped some levels of variables that had few samples. Age was

discretized into 3 categories ((15, 40], (40, 60], and (60, 90] years), IUCC stages were regrouped into four categories (I, II, III, IV), and smoking status was regrouped into 2 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 Figure 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, Nature, (2015)[14] displayed Y chromosome expression patterns discordant with their clinical data (S02249 and S02293; **Supplementary Figure 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 Table 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 Figure 29A**). Indeed, the 20 carcinoids from one of the providers (provider 1) are all atypical carcinoids. Nevertheless, because there are also 7 atypical carcinoids from a second provider and 5 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 Figure 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[4,11,12,14,15]. For the new ones, an HE (hematoxylin 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 (EB, BAA, and SL). 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 (3µm thick) from twenty 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 tumor 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 more than 5%. Results are presented **Supplementary Table 9** and representative slides are displayed in **Figure 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 to survival analyses is provided in **Supplementary Table 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, Nature (2015)[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 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 15min at 30ºC. 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).

DNA Sequencing

Whole-Genome sequencing (WGS)

Whole-genome sequencing was performed on 3 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 (1µg) 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 file 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 ~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 primer 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 6 separated primers pools ((1)1 pool covering 786 amplicons, (2) 4 pools covering 498 amplicons, and (3) 1 pool covering 47 amplicons). Per pool, 20ng (1) or 10ng (2 and 3) of DNA were dispensed and air-dried (only 2 and 3). Subsequently 11µL (1) or 5µl (2 and 3) of the PCR mix were added [containing 5.5µL (1) or 2.5µl (2 and 3) Primer mix pool (2X), 2.2µL (1) or 1µl (2 and 3) PCR Buffer (5X), 0.73µL (1) or 0.34µl (2 and 3) HotStar Taq DNA Polymerase (6U/µl) and 0.57µL (1) or 1.16µl (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 sec 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). 100ng of purified PCR product (6µl) 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 [1.5µl of NEBNext End Repair Reaction Buffer, 0.75µL of NEBNext End Repair Enzyme Mix, and 6.75µL of H2O] followed by ligation to specific adapters and in-house prepared individual barcodes (Eurofins MWG Operon, Germany) [4.35µl of H2O, 2.5µl of T4 DNA Ligase Buffer for Ion Torrent, 0.7µl of Ion P1 adaptor (double-stranded), 0.25µl of Bst 2.0 WarmStart DNA Polymerase, 1.5µl of T4 DNA ligase, and 0.7µl of in-house barcodes]. Bead purification of 1.8X was applied to clean libraries and 100 ng of adaptator ligated DNA were amplified with 15µl of Master Mix Amplification [containing 1µl of Primers, 12.5µl of NEBNext High-Fidelity 2X PCR Master Mix, and 1.5µl of H2O]. Pooling of libraries was performed equimolarly and loaded on a 2% agarose gel for electrophoresis (220V, 40 minutes). 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 manufacture's protocols.

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 in 3 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 Figure 28A**).

Variant calling and filtering

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 characterization 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 analyzed 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 (10 typical and 5 atypical carcinoids) previously analyzed(Fernandez-Cuesta et al, Nat Comm 2013)[11] were reanalyzed 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 tumors). 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 tumor 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 tumor.

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 Table 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 in three main steps. First, the dependency between GC content and raw read count is modeled using a generalized 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 **Supplementary Table 8** and one sample is highlighted in **Figure 3C**.

Gene-set enrichment analysis

Gene-set enrichment for somatic mutations was assessed independently for each set of Hallmark of cancer genes[19] using a two-sided Fisher's exact test, and followed by a correction for multiple testing[46]. 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 Table 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 Table 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 Table 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 Table 5.**

RNA Sequencing

RNA-seq data

RNA sequencing was performed on 20 fresh frozen atypical carcinoids in the Cologne Center for Genomics. Libraries were prepared using the Illumina® TruSeq® RNA sample preparation Kit. Library preparation started with 1µg 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 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.

Data processing

The 210 raw reads files (89 carcinoids, 69 LCNEC, 52 SCLC) were processed in 3 steps using the RNA-seq processing workflow based on the nextflow language[47] and accessible at https://github.com/IARCbioinfo/RNAseq-nf (revision da7240d). Reads were scanned for a part of Illumina's 13bp adapter sequence 'AGATCGGAAGAGC' at the 3' end using Trim Galore v0.4.2

(Krueger 2015) with default parameters. 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. (2016)[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-stabilized read counts (vst function from R package DESeq2) of each sample on the X and Y chromosomes (**Supplementary Figure 28B**).

Variant calling

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 analysis

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 Table 7**, and selected genes are highlighted in **Figure 3B.**

Unsupervised analysis of expression data

The raw read counts of 57,822 genes from the 210 samples were normalized using the variance stabilization 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-stabilized 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 Figure 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 dataset. We did not detect any outliers in any of the analyses from **Supplementary Figure 6**. We further studied the association between expression data, batch (sample provider), and 5 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 6 covariables of interest (provider plus the 5 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 Figure 30A**). The fact that both histopathology and sample provider are jointly significantly associated with PC2 is expected given their non-independence (**Supplementary Figure 29A** and **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 Figure 30B**). We can see that samples from Provider 1 and provider 2 span a similar range on PC2 (from values lower 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 (5 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 Figure 30B**).

Supervised analysis of expression data

We performed three distinct differential expression (DE) analyzes. i) A comparison between histopathological types; ii) A comparison between pulmonary carcinoid (LNET) clusters A1, A2, and B (see **Figure 5A** and the Multi-omic integration method section); iii) a comparison between

lung neuro-endocrine 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 analyzes (samples S01093, S02236), and further excluding three samples with no clear histopathological type (classified as carcinoids in **Supplementary Table 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 (2 levels: male and female), and age (3 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 Table 15** and **Supplementary Figure 31**; results of analysis (ii) are reported in **Supplementary Table 10** and **Figure 5A**; results of analysis (iii) are reported in **Supplementary Table 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 behavior 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 Figure 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 10 immune cell types (B cells, M1, 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 March 23 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 dataset containing expected expression levels for the 10 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 3 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 **Figures 2F and 4B, and Supplementary Figures 15, 19 and 32**.

Methylome analyses

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 8 samples, so 12 chips were needed for the 95 samples. We used stratified randomization 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 bi-sulfite converted using the EZ-96 DNA Methylation-Gold™ kit (Zymo Research Corp., CA, USA) following the manufacturer's recommendations for Infinium assays. 3 replicates included half the amount (300ng). Then, 200 ng of bisulfite-converted DNA was used for hybridization on Infinium MethylationEPIC

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).

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 normalization 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-hybridize 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 detectionP 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 minimize potential discrepancies, and we removed one sample that came from a metastatic tumour rather than a primary tumour (sample not reported in the study). The final dataset 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, hybridization). 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 Figure 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 10 principal components—hereafter referred to as surrogate variables—explaining more than 90% of the variation in control probes intensity. The 10 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 Figure 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 (8483 and 7693 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 Figure 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 dataset. We did not detect any outliers in any of the analyses from **Supplementary Figure 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 Figure 33A**). Principal component 2 was associated with the sample provider; further examination of the PCA (**Supplementary Figure 33B**) revealed that this effect was driven by the samples from provider 1, which have the largest range of coordinates on PC2 (from less than -30 to more than 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 normalization on batch effects by performing the same analysis on the *M*-values of the 5% most variable CpGs obtained without normalization (**Supplementary Figure 33A**). Compared to the PCA of the 5% most variable CpGs with normalization (**Supplementary Figure 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 normalization 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 Figure 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 10 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 favored in array analyses. The *p*-values were adjusted for multiple testing, and CpGs with a *q*-value lower than 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 Tables 16**, **Supplementary Tables 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 **Figure 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

Data

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

Unsupervised analyses

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)[16]. MOFA identifies latent factors (LF, i.e., continuous variables) that explain most variation in the joint datasets. We did not include the somatic mutations in the model because the low level of recurrence (only 4 recurrently-mutated genes in **Supplementary Table 4**) resulted in a sample by mutation matrix of much lower dimension than the other 'omics, which is known to bias the analyses[16]. 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 sections RNA sequencing and EPIC 850k methylation array). 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 datasets (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 less than 2% of the variance of both 'omic datasets (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 5 first latent factors across each run, resulting in a correlation matrix of 100 by 100 entries (**Supplementary Figures 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 datasets. 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 **Figures 1A, 4A, and Supplementary Figure 13**. Interestingly, we find that MOFA latent factors 1 to 3 for analysis (i) (LNET, LCNEC, and SCLC) correspond to MOFA LF 2 to 4 for analysis (ii) (LNET and LCNEC), and to MOFA LF 3 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**.**

Comparison with uni-omic unsupervised analyses. We compared the results of MOFA with that of the unsupervised analysis of expression and methylation data (**Supplementary Figure 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 Figure 3A**), we found that MOFA LF1 is strongly correlated with expression PC1 and methylation PC1 ($|r|>0.98$; **Supplementary Figure 3D** and **E**), and that expression PC1 and methylation PC1 are strongly correlated between them ($r=0.97$; **Supplementary Figure 3C**); LF2 was strongly correlated with expression PC3 ($r=-0.86$; **Supplementary Figure 3P**), and methylation PC2 ($r=-0.98$; **Supplementary Figure 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 Figure 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 (**Figure 1A**); it is also coherent with the fact that clusters A1 and A2 are the most separated clusters on expression PC2 (**Supplementary Figure 6B**), while clusters A1 and B are the most separated on methylation PC2 (**Supplementary Figure 7A**). When using LNET samples only (**Supplementary Figure 3B**), we found that MOFA LF1 is strongly correlated with expression PC2 and methylation PC1 ($|r|>0.86$; **Supplementary Figure 3M** and **H**), and that expression PC2 and methylation PC1 are strongly correlated between them ($r=0.72$; **Supplementary Figure 3F**); LF2 was strongly correlated with expression PC1 ($r=-0.88$; **Supplementary Figure 3G**), and methylation PC2 ($r=0.90$; **Supplementary Figure 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 (**Figure 4A**); it is also coherent with the fact that clusters A1 and A2 are the most separated clusters on expression PC1 (**Supplementary Figure 6D**), while clusters A1 and B are the most separated on methylation PC2 (**Supplementary Figure 7B**).

Associations of latent factors with other variables. We used the results from MOFA to detect outliers and batch effects in the dataset. We did not detect any outliers in any of the analyses from **Supplementary Figure 13**. We further studied the associations between the first 5 LFs, batch (sample provider), and 5 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 6 covariables of interest (provider plus the 5 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 3 samples (following Argelaguet *et al*. 2017)[16]; 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 **Figure 4A**. Similarly, we tested the association between pathways highlighted in **Supplementary Figure 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.

Clustering. We identified molecular clusters—groups of samples with similar molecular profiles—from MOFA results. Following Mo et al, PNAS, (2013)[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, PNAS [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 dataset—, 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 Bioinformatics (2010)[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 **Supplementary Figures 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 Figure 18C**), we chose the solution with 3 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 Figure 18B** are split between a blue and a purple cluster for *K*=4), so the solutions with 3 and 4 clusters are coherent. Cluster memberships are highlighted **Figure 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 5 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 minimizing the within-cluster sum of squares, the weighted within-cluster sum of squares is minimized. Results for *K*=3 clusters of LNET and LNEN samples are presented in **Supplementary Figure 8**. We can see that the alternative approach (weighted *K*-means on 5 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 = 4.09×10^{-84}, 8.63×10^{-80}, 0.66, 0.094, 0.24, respectively for latent factors 1 through 5) or the LNET clusters (ANOVA q = 5.06×10^{-4}, 5.99×10^{-47}, 5.12×10^{-46}, 0.15, 0.052, respectively), which indicates that the first 3 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 of features associated with latent factors. We performed Gene set enrichment analysis (GSEA) on the latent factors identified by MOFA using the built-in function FeatureSetEnrichmentAnalysis [16]. 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 muti-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 **Supplementary Table 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 2.2.0based on the IlluminaHumanMethylationEPICanno.ilm10b2.hg19 annotation (getAnnotation 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 Tables 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 linear model independently for each correlated CpGs (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 higher than 0.25 were retained and, among these CpGs, only the one with the smallest *q*-value has been represented in **Supplementary Figure 22**. Results of analyses (i) and (ii) are reported in **Supplementary Tables 10** and **12**.

Survival analysis using penalized generalized linear model.

We computed a generalized linear model with elastic net regularization (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 regularization parameter λ (cv.glmnet function from glmnet package). To be more stringent, the optimal regularization parameter chosen was the one associated with the most regularized 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 normalized read counts, and centered and scaled them using R package caret (v6.0-80). The genes with non-zero estimated coefficients are listed in **Supplementary Table 13**. For each non-coding gene, we determined the optimal cutpoint of expression (normalized 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 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 normalization 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 normalized using the variance stabilization transformation (vst function from R package DESeq2 v1.22.2), then

mean-centered and scaled to unit variance. Then, the variance stabilizing transformation learned from the training set was applied to the test set using the dispersionFunction function from the DESeq2 package, and centering 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-centered and scaled to unit variance, then the test sample features were centered 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 Table 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 Figure 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 Figure 31C, left panel**), the only feature considered was the expression of *MKI67*.

Integration of expression and methylation data. Because 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 Figure 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 dataset separately corresponds to maximizing 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 dataset. An alternative approach is to maximize the number of features (p) by combining both datasets, 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 maximize n. In addition, by maximizing n, we hypothetically ensured that we would also maximize 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 Figure 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 **Figure 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 Figure 9**). This supports our hypothesis that maximizing 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.

Because matrix factorization methods such as MOFA and PCA remove correlations between features by finding latent factors that summarize 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 (5 MOFA latent factors, 6 expression PCs, and 5 methylation PCs, respectively). These analyses are presented in **Supplementary Figure 12** and led to similar classification to the results presented in the main text **Figure 1**. In addition, in the case of MOFA factors, in accordance with **Figure 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. We first performed a survival analysis using the predictions 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 more than 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 Figure 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 Figure 11C**).

Comparison between the supervised analyses. 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 Figure 31**). Survival analyses showed a significant difference between histopathological types (**Supplementary Figure 31A**). Nevertheless, the machine learning classifier based on the genome-wide expression or methylation data could not properly distinguish atypical and typical carcinoids (**Supplementary Figure 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 Figure 31C**,middle panel; **Supplementary Table 15**) and differentially methylated positions (**Supplementary Figure 31C**, right panel; **Supplementary Table 17**). Overall, these data suggest that the histopathological classification, although clinically meaningful, does not completely match the molecular classification.

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 EGAS#1096. Other datasets 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 summarized in **Supplementary Table 18.**

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. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **10**, 1243-1260, doi:10.1097/jto.0000000000000630 (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 pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, doi:10.1038/s41379-018-0110-y (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. *Annals of oncology : official journal of the European Society for Medical Oncology* **26**, 1604-1620, doi:10.1093/annonc/mdv041 (2015).

4 Swarts, D. R. *et al.* Interobserver variability for the WHO classification of pulmonary carcinoids. *The American journal of surgical pathology* **38**, 1429-1436, doi:10.1097/pas.0000000000000300 (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. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **12**, 334-346, doi:10.1016/j.jtho.2016.12.004 (2017).

6 Marchio, C. *et al.* Distinctive pathological and clinical features of lung carcinoids with high proliferation index. *Virchows Archiv : an international journal of pathology* **471**, 713-720, doi:10.1007/s00428-017-2177-0 (2017).

7 Pelosi, G., Rindi, G., Travis, W. D. & Papotti, M. Ki-67 antigen in lung neuroendocrine tumors: unraveling a role in clinical practice. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **9**, 273-284, doi:10.1097/jto.0000000000000092 (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. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **13**, 752-766, doi:10.1016/j.jtho.2018.02.002 (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 Archiv : an international journal of pathology* **472**, 567-577, doi:10.1007/s00428-018-2307-3 (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. *Clinical cancer research : an official journal of the American Association for Cancer Research* **22**, 3618-3629, doi:10.1158/1078-0432.ccr-15-2946 (2016).

11 Fernandez-Cuesta, L. *et al.* Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nature communications* **5**, 3518, doi:10.1038/ncomms4518 (2014).

12 George, J. *et al.* Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nature communications* **9**, 1048, doi:10.1038/s41467-018-03099-x (2018).

13 Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature genetics* **44**, 1104-1110, doi:10.1038/ng.2396 (2012).

14 George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47-53, doi:10.1038/nature14664 (2015).

15 Swarts, D. R. *et al.* CD44 and OTP are strong prognostic markers for pulmonary carcinoids. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 2197-2207, doi:10.1158/1078-0432.ccr-12-3078 (2013).

16 Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology* **14**, e8124, doi:10.15252/msb.20178124 (2018).

17 IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. in *IARC Monographs on the evaluation of carcinogenic risks to humans: Arsenic, Metals, Fibres and Dusts* Vol. 100C 219-309 (2012).

18 Carbone, M. *et al.* BAP1 and cancer. *Nature reviews. Cancer* **13**, 153-159, doi:10.1038/nrc3459 (2013).

19 Kiefer, J. *et al.* Abstract 3589: A systematic approach toward gene annotation of the hallmarks of cancer. *Cancer research* **77**, 3589-3589, doi:10.1158/1538-7445.am2017-3589 (2017).

20 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

21 Shi, C. & Pamer, E. G. Monocyte recruitment during infection and inflammation. *Nature reviews. Immunology* **11**, 762-774, doi:10.1038/nri3070 (2011).

22 Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. *Nature reviews. Immunology* **13**, 159-175, doi:10.1038/nri3399 (2013).

23 Jakubzick, C. V., Randolph, G. J. & Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nature reviews. Immunology* **17**, 349-362, doi:10.1038/nri.2017.28 (2017).

24 Cernadas, M., Lu, J., Watts, G. & Brenner, M. B. CD1a expression defines an interleukin-12 producing population of human dendritic cells. *Clinical and experimental immunology* **155**, 523-533, doi:10.1111/j.1365-2249.2008.03853.x (2009).

25 Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science (New York, N.Y.)* **303**, 1378-1381, doi:10.1126/science.1089769 (2004).

26 Tran Janco, J. M., Lamichhane, P., Karyampudi, L. & Knutson, K. L. Tumor-infiltrating dendritic cells in cancer pathogenesis. *Journal of immunology (Baltimore, Md. : 1950)* **194**, 2985-2991, doi:10.4049/jimmunol.1403134 (2015).

27 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. *Nature reviews. Cancer* **17**, 765, doi:10.1038/nrc.2017.106 (2017).

28 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. *The Lancet. Oncology* **18**, 42-51, doi:10.1016/s1470-2045(16)30565-4 (2017).

29 Gara, R. K. *et al.* Slit/Robo pathway: a promising therapeutic target for cancer. *Drug discovery today* **20**, 156-164, doi:10.1016/j.drudis.2014.09.008 (2015).

30 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. *American journal of respiratory and critical care medicine* **154**, 758-763, doi:10.1164/ajrccm.154.3.8810616 (1996).

31 Sutherland, K. D. & Berns, A. Cell of origin of lung cancer. *Molecular oncology* **4**, 397-403, doi:10.1016/j.molonc.2010.05.002 (2010).

32 Branchfield, K. *et al.* Pulmonary neuroendocrine cells function as airway sensors to control lung immune response. *Science (New York, N.Y.)* **351**, 707-710, doi:10.1126/science.aad7969 (2016).

33 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 immunology, immunotherapy : CII* **64**, 51-59, doi:10.1007/s00262-014-1613-0 (2015).

34 Scarpa, A. *et al.* Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature* **543**, 65-71, doi:10.1038/nature21063 (2017).

35 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. *The Journal of pathology* **241**, 488-500, doi:10.1002/path.4853 (2017).

36 Papaxoinis, G. *et al.* Prognostic Significance of CD44 and Orthopedia Homeobox Protein (OTP) Expression in Pulmonary Carcinoid Tumours. *Endocrine pathology* **28**, 60-70, doi:10.1007/s12022-016-9459-y (2017).

37 Koyama, T. *et al.* ANGPTL3 is a novel biomarker as it activates ERK/MAPK pathway in oral cancer. *Cancer medicine* **4**, 759-769, doi:10.1002/cam4.418 (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, doi:10.1038/onc.2015.185 (2016).

39 Williams, C. S. *et al.* ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. *Carcinogenesis* **36**, 710-718, doi:10.1093/carcin/bgv049 (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 Archiv : an international journal of pathology* **471**, 31-47, doi:10.1007/s00428-017-2130-2 (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, doi:10.1136/gutjnl-2016-313075 (2017).

42 Mojic, M., Takeda, K. & Hayakawa, Y. The Dark Side of IFN-gamma: Its Role in Promoting Cancer Immunoevasion. *International journal of molecular sciences* **19**, doi:10.3390/ijms19010089 (2017).

43 Zaidi, M. R. & Merlino, G. The two faces of interferon-gamma in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 6118-6124, doi:10.1158/1078-0432.ccr-11-0482 (2011).

44 Ocana, A., Nieto-Jimenez, C., Pandiella, A. & Templeton, A. J. Neutrophils in cancer: prognostic role and therapeutic strategies. *Molecular cancer* **16**, 137, doi:10.1186/s12943-017-0707-7 (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. *The American journal of surgical pathology* **40**, 1192-1202, doi:10.1097/pas.0000000000000662 (2016).

46 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

47 Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35**, 316-319, doi:10.1038/nbt.3820 (2017).

48 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

49 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics (Oxford, England)* **30**, 2503-2505, doi:10.1093/bioinformatics/btu314 (2014).

50 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)* **31**, 2032-2034, doi:10.1093/bioinformatics/btv098 (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 (Oxford, England)* **30**, 2813-2815, doi:10.1093/bioinformatics/btu376 (2014).

52 Delhomme, T. M. *et al.* needlestack: an ultra-sensitive variant caller for multi-sample deep next generation sequencing data. *bioRxiv* (2019). <Preprint at >.

53 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

54 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

55 Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Current protocols in bioinformatics* **51**, 11.14.11-19, doi: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 (Oxford, England)* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (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. *Nature protocols* **11**, 1650-1667, doi:10.1038/nprot.2016.095 (2016).
58 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).
59 Fernandez-Cuesta, L. *et al.* Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome biology* **16**, 7, doi:10.1186/s13059-014-0558-0 (2015).
60 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
61 Dray, S. & Dufour, A. B. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of statistical software* **22**, 20, doi:10.18637/jss.v022.i04 (2007).
62 Finotello, F. *et al.* quanTIseq: quantifying immune contexture of human tumors. *bioRxiv* https://doi.org/10.1101/223180 (2017).
63 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).
64 Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic acids research* **44**, e20, doi:10.1093/nar/gkv907 (2016).
65 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47, doi:10.1093/nar/gkv007 (2015).
66 Mo, Q. *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 4245-4250, doi:10.1073/pnas.1208949110 (2013).
67 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics (Oxford, England)* **26**, 1572-1573, doi:10.1093/bioinformatics/btq170 (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. *Nature protocols* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).
69 Zhang, J. D. & Wiemann, S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics (Oxford, England)* **25**, 1470-1471, doi:10.1093/bioinformatics/btp167 (2009).
70 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **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.

Funding

This study is part of the lungNENomics project. This work has been funded by the US National Institutes of Health (NIH R03CA195253 to LFC and JDM), the French National Cancer Institute (INCa, PRT-K-17-047 to LFC and TABAC 17-022 to JDM), the Ligue Nationale contre le Cancer (LNCC 2016 to LFC), France Genomique (to JDM), and the Italian Association for Cancer Research

(AIRC) (IG 19238 to MV) (Special Program 5X1000, ED No12162 to UP, LR and GS). JS is a Miguel Servet researcher (CP13/00055 and PI16/0295). LM and TMD have fellowships from the LNCC.

Author contributions

LFC conceived and designed the study. LFC and MF supervised all the aspects of the study. AG, AB, JA, FLCK, SB, JS, NG, and SL supervised some aspects of the study. BAA, EB, and SL performed the histopathological review. NoL, TG, JD, AC, CCu, GD, and NiL did the lab work. NA, NoL, AAGG, LM, DH, ASS, AF, TMD, RO, VM, CV, and LaM performed the computational and statistical analyses. PL, ACT, AML, AS, JHC, JSa, JSt, JKF, MB, CBF, FGS, NLS, PAR, GW, LR, GS, UP, MM, SL, JMV, VH, PH, OTB, MLI, VTM, LAM, PG, MV, MGP, LB, HP, AMCD, EB, EJMS, NG, and SyL contributed with samples and the corresponding histopathological, epidemiological, and clinical data. JFD, ZH, AV, PN, and JDM helped with logistics. JD, BAA, CCa, LR, MM, MV, MGP, LB, HP, GP, JDM, HHV, EJMS, NG, and SL gave scientific input. NA, NoL, AAGG, LM, JDM, MF, and LFC wrote the manuscript, which was reviewed and commented by all the co-authors.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 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 Panel 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 Panel 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. Source data are provided in **Supplementary Table 1**.

Figure 2. Molecular characterization of supra-carcinoids. A) Forest plot of hazard ratios for overall survival of the supra-carcinoids, compared to carcinoid clusters A and B, and LCNEC. The number of samples (N) in each group is given in brackets. The back 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. Source data are provided in **Supplementary Tables 1, 4, 5, 12, 17**, and in the European Genome-phenome Archive.

Figure 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. **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. Source data are provided in **Supplementary Tables 4**, **7**, and **8**.

Figure 4. Multi-omics unsupervised analysis of lung neuroendocrine tumours. A) Multi-Omics Factor Analysis (MOFA) of transcriptomes and methylomes of restricted to LNET samples (pulmonary carcinoids). Design follow that of **Figure 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 (HE) 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.. Source data are provided in **Supplementary Tables 1, 4, 9,** and in the European Genome-phenome Archive.

Figure 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 color scale depends on the range of q-value (q) and correlation estimate squared (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 regularized 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. Cut-offs 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. Source data are provided in **Supplementary Tables 5, 10,** and in the European Genome-phenome Archive.

Figure 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. Source data are provided in the European Genome-phenome Archive.

Tables

	LNEN005	LNEN012	LNEN021	LNEN022	S01513	S01522
CLASSIFICATION						
Histopathology	Atypical	Atypical	Atypical	Atypical	Atypical	Atypical
Morphological characteristics	carcinoid morph. 2 mitoses/2mm^2 No necrosis	carcinoid morph. 2 mitoses/2mm^2 No necrosis	LCNEC morph. 4 mitoses/2mm^2 No necrosis	NA	NA	NA
Machine learning	LCNEC	LCNEC	Unclassified	Unclassified	Atypical	Unclassified
CLINICAL DATA						
Sex	M	F	F	F	M	M
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*	8.12	10.32	NA	NA	3.15	NA
MKI67 FPKM	2.6	7.3	NA	NA	1.9	NA

* IC genes median FPKM values for pulmonary carcinoids, LCNEC and SCLC are 1.0, 3.5, and 3.2, respectively

Table 1. Characteristics of supra-carcinoids. FPKM refers to Fragments Per Kilo-base per Million reads. The median FPKM of immune checkpoint (IC) genes was calculated based on the genes included in **Figure 2E**, excluding HLA genes because of their very large expression levels

List of Supplementary Figures

Supplementary Figure 1 Overview of the multi-omics experimental design for LNEN samples

Supplementary Figure 2 Robustness of the MOFA latent factors presented in Figure 1A

Supplementary Figure 3 Correlations between MOFA latent factors (Figs. 1A and 4A) and the principal components of the PCA of expression (Supplementary Figure 6) and methylation (Supplementary Figure 7)

Supplementary Figure 4 Forest plot of the survival analysis based on the first three MOFA latent factors (LFs) of LNEN samples from Figure 1A

Supplementary Figure 5 Robustness of the consensus clustering of LNENs presented in Figure 1A

Supplementary Figure 6 Principal Component Analysis (PCA) of transcriptome data

Supplementary Figure 7 Principal Component Analysis of the methylation data

Supplementary Figure 8 Comparison between consensus clustering on MOFA latent factors based on different clustering algorithms

Supplementary Figure 9 Analysis of the ML predictions based on a model integrating expression and methylation data simultaneously

Supplementary Figure 10 Comparison of the ML predictions when applying different thresholds to define the Unclassified category

Supplementary Figure 11 Comparison of overall survival based on different classifications

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

Supplementary Figure 13 Consistency of MOFA across analyses including different histopathological types

Supplementary Figure 14 Radar chart of the expression levels of HLA class I and related immunostimulatory genes as a function of their molecular group

Supplementary Figure 15 Estimation of the amount immune cells in the different pulmonary carcinoid groups from transcriptome data

Supplementary Figure 16 Cancer-relevant somatically altered pathways altered in typical and atypical carcinoids

Supplementary Figure 17 Robustness of the MOFA latent factors presented in Figure 4A

Supplementary Figure 18 Robustness of the consensus clustering of pulmonary carcinoids presented in Figure 4A

Supplementary Figure 19 Estimation of the amount of immune cells in the different LNET clusters and supra-carcinoids from transcriptome data

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

Supplementary Figure 21 Comparison of two methods to identify core differentially expressed (DE) genes of LNET clusters

Supplementary Figure 22 Correlations between DNA methylation and gene expression for core genes of LNET clusters

Supplementary Figure 23 DNA methylation and gene expression levels of *HNF1A* and *HNF4A* in LNET samples

Supplementary Figure 24 Expression levels of NOTCH genes in the different LNET clusters, supra-ca, LCNEC and SCLC

Supplementary Figure 25 Correlation between DLL3 and CDA1 expression based on immunohistochemistry in a validation series

Supplementary Figure 26 Survival (Kaplan-Meier curve) of *MEN1* wild type compared to mutant cases

Supplementary Figure 27 Expression levels of core cluster B genes associated with survival (Figure 1B)

List of Supplementary Tables