Molecular characterisation of pulmonary supra-carcinoids through integration of multi-omics data

Alexandra Sexton-Oates, Alex Di Genova, Nicolas Alcala, Lise Mangiante, Catherine Voegele, Matthieu Foll, Lynnette Fernandez-Cuesta, and the lungNENomics team

Objectives

Pulmonary carcinoids (PCA), comprising typical and atypical carcinoids, are a group of low grade rare lung neuroendocrine tumours that, unlike their high grade counterparts, have relatively good prognosis and no known risk factors. Despite their better prognosis, many PCA patients present with metastatic disease or suffer relapse, both of which respond poorly to current therapeutic regimens, and identifying patients at greatest risk of relapse is difficult. Furthermore, little is known about the recently described supra-carcinoid subtype, which are morphologically similar to atypical carcinoids, yet more aggressive. Given their rarity, characterisation of these tumours has been limited, particularly with regard to whole-genome sequencing of atypical and supra-carcinoid tumours. In the lungNENomics project, part of the Rare Cancers Genomics initiative, we aim to perform comprehensive multi-omic molecular, morphological, and clinical characterisation of PCA, including the supra-carcinoid subtype. This study will help us to understand the biological mechanisms underlying the development of PCA, and improve characterisation and classification of supra-carcinoids.

Methods

We have generated whole-genome sequencing (WGS), RNA sequencing and DNA methylation array data for tumour specimens from 91 patients, including 58 patients with atypical tumour type. These data have been combined with previously published RNA sequencing and DNA methylation array data for predominantly typical carcinoids, and higher grade lung neuroendocrine tumours, in order to perform integrative multi-omics factor analysis for molecular characterisation (Argelaguet *et al.* Mol Syst Biol 2018).

Results

Previously the Rare Cancers Genomics team have published an analysis of gene expression and DNA methylation array data for a collection of PCA, uncovering clinically relevant groups and the existence of the new supra-carcinoid entity (Alcala *et al*. Nat Comms 2019). In the current study we have improved upon previous efforts by incorporating a large number of new samples, enriched for atypical carcinoids, the majority of which have also been subjected to WGS. These data have been integrated to expand our first molecular map of PCA (Gabriel *et al.* GigaScience 2020), identify further instances of rare supra-carcinoids, and characterise the molecular biology of PCA types. Finally, tumour map positions were used to perform archetype analysis (Hart *et al.* Nat Methods 2015) to uncover evolutionary trade-offs in cancer-specific tasks between tumour types.

Conclusions

While progress has been made in recent years in the molecular characterisation of PCA, there are many clinically-relevant questions which remain. These can be addressed by investigating all molecular layers, including whole-genome sequencing, that until now has been lacking in PCA. The lungNENomics project aims to address these important questions, and to improve the biological understanding of tumour development and progression in this exceedingly rare and understudied disease.

References

1. Alcala N et al. Nature Communications, 2019 PMID:31431620

2. Simbolo M et al. Journal of Thoracic Oncology, 2019 PMID: 31085341

3. Fernandez-Cuesta L et al. Nature Communications, 2014 PMID: 24670920

4. George J et al. Nature Communications, 2018 PMID: 29535388

5. Laddha SV et al. Cancer Research, 2019 PMID: 31300474

6. Miyanaga A et al. Lung Cancer, 2020 PMID: 32417679

7. Argelaguet R et al. Molecular Systems Biology, 2018 PMID: 29925568

8. Gabriel A et al. GigaScience, 2020 PMID: 33124659

nature COMMUNICATIONS

ARTICLE

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

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

N. Alcala et al.#

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

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

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

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

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

Results

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

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

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

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

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

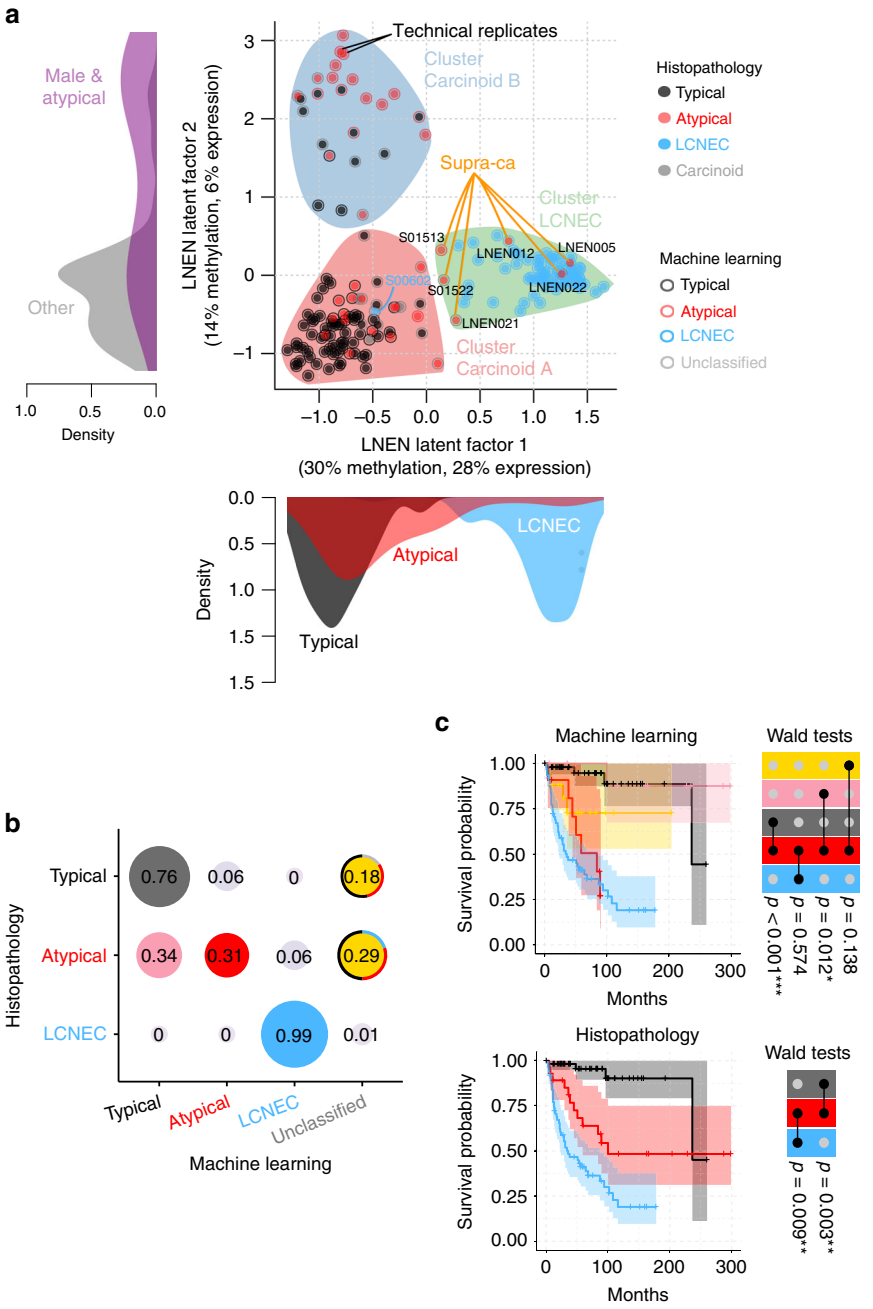


Fig. 1 Multi-omics (un)supervised analyses of lung neuroendocrine neoplasms. **a** Multi-omics factor analysis (MOFA) of transcriptomes and methylomes of LNEN samples (typical carcinoids, atypical carcinoids, and LCNEC). Point colours correspond to the histopathological types; coloured circles correspond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

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

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

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

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

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

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

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

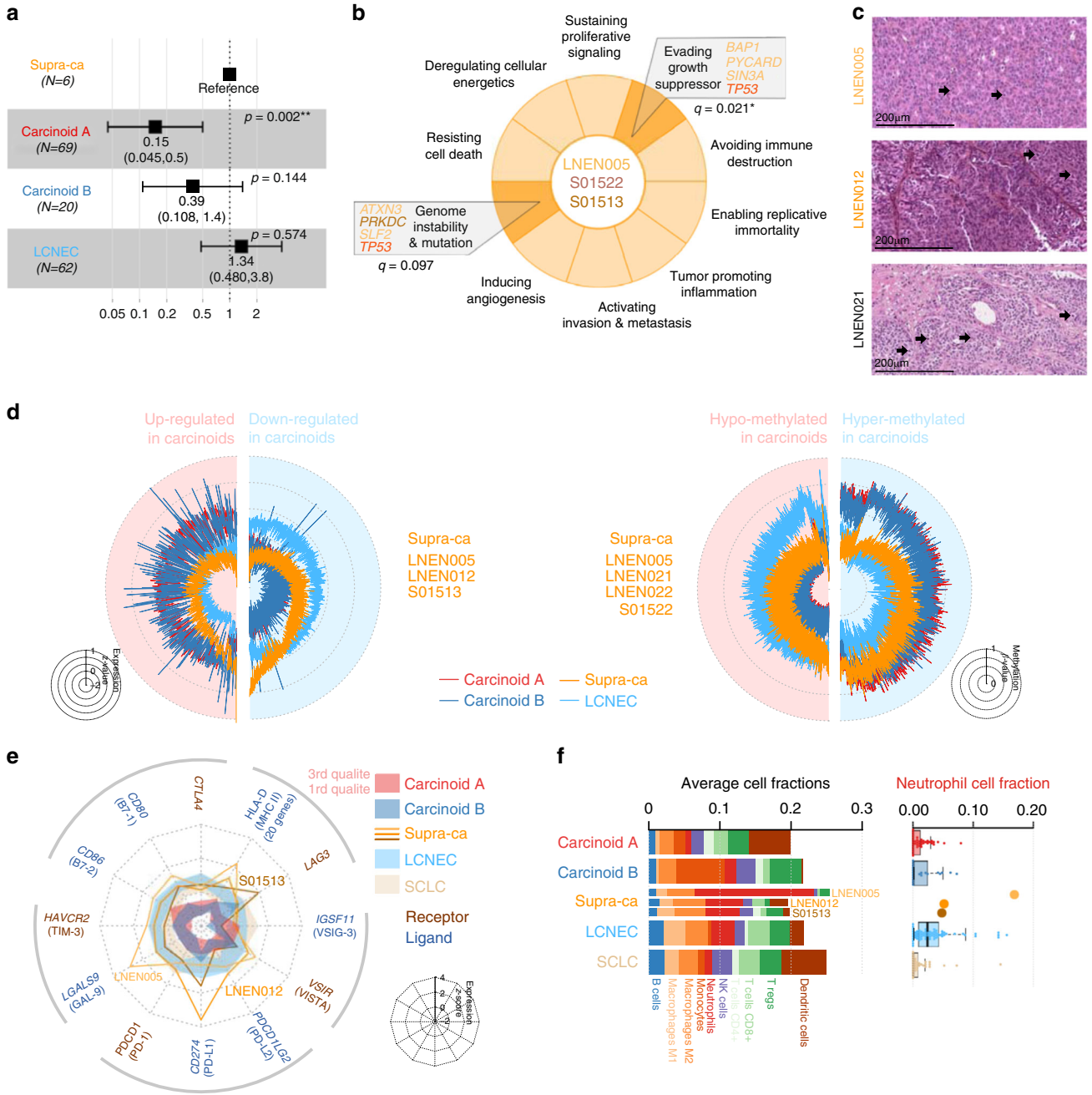


Fig. 2 Molecular characterisation of supra-carcinoids. **a** Forest plot of hazard ratios for overall survival of the supra-carcinoids, compared to Carcinoid A and B, and LCNEC. The number of samples (*N*) in each group is given in brackets. The black box represent estimated hazard ratios and whiskers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

Table 1 Characteristics of supra-carcinoids

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

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

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

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

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

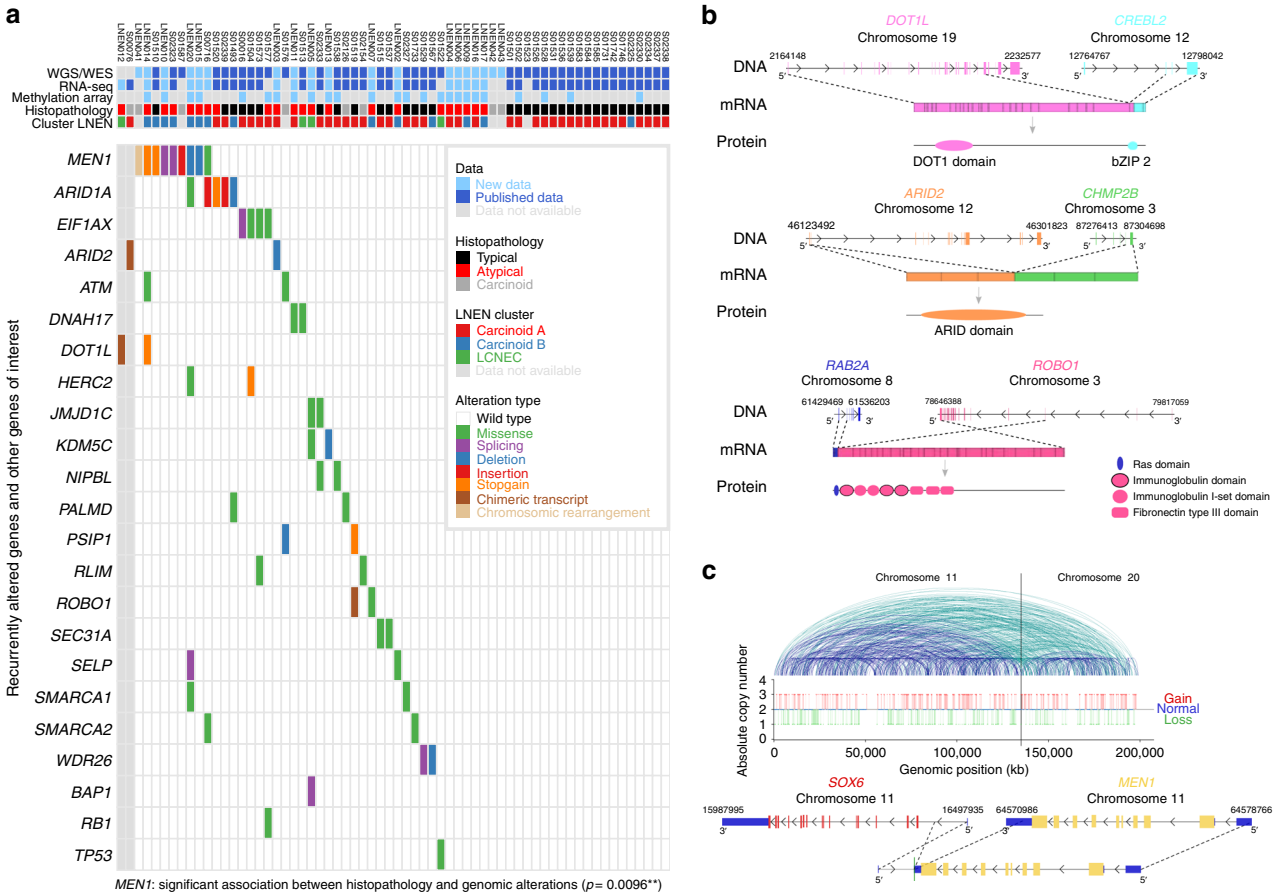


Fig. 3 Mutational patterns of pulmonary carcinoids. **a** Recurrent and cancer-relevant altered genes found in pulmonary carcinoids by WGS and WES. Fisher's exact test *p*-value for the association between *MEN1* and the atypical carcinoid histopathological subtype is given in brackets; $0.01 \le p 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

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

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

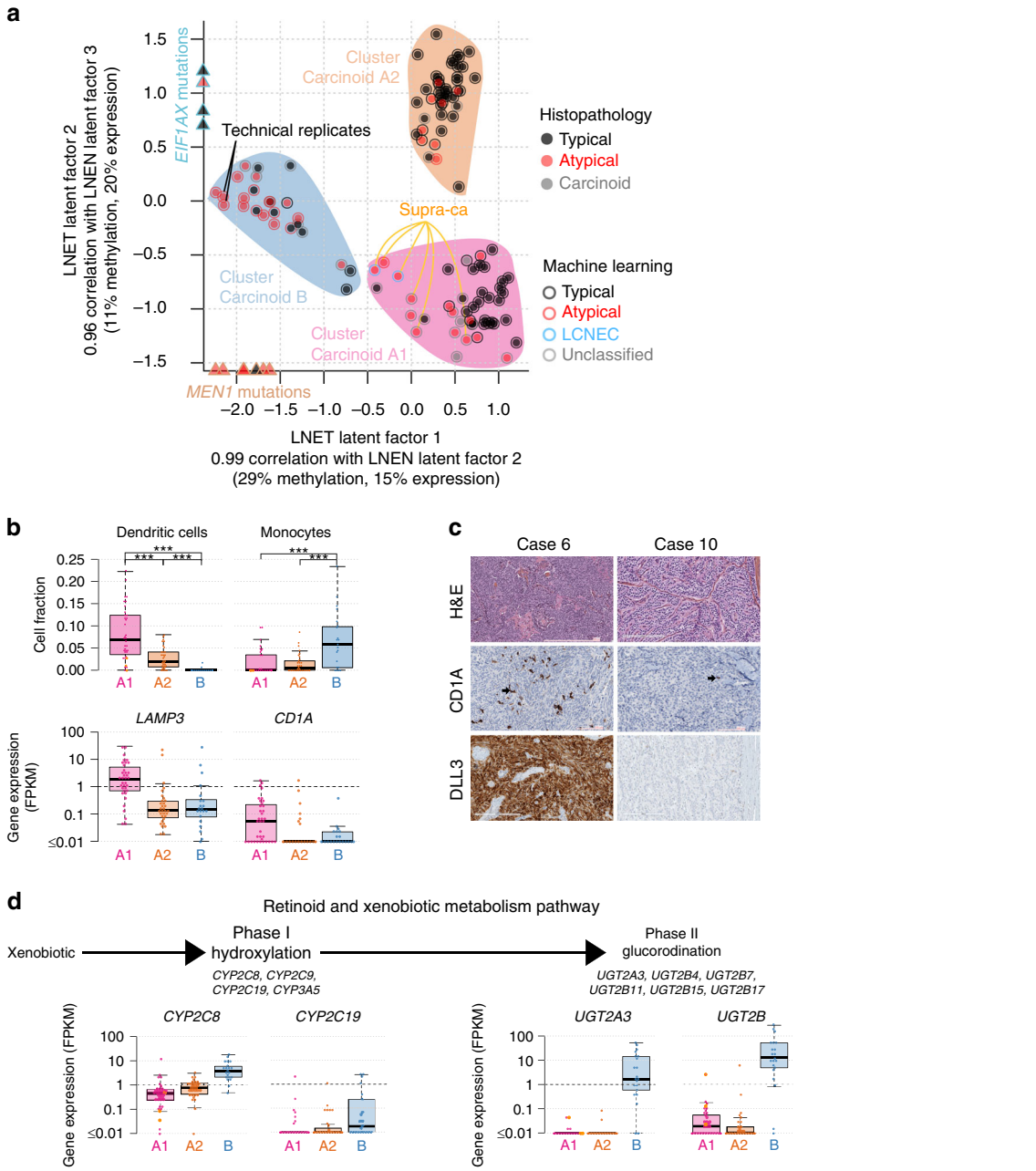


Fig. 4 Multi-omics unsupervised analysis of lung neuroendocrine tumours. **a** Multi-omics factor analysis (MOFA) of transcriptomes and methylomes restricted to LNET samples (pulmonary carcinoids). Design follows that of Fig. 1a; filled coloured shapes represent the three molecular clusters (Carcinoid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

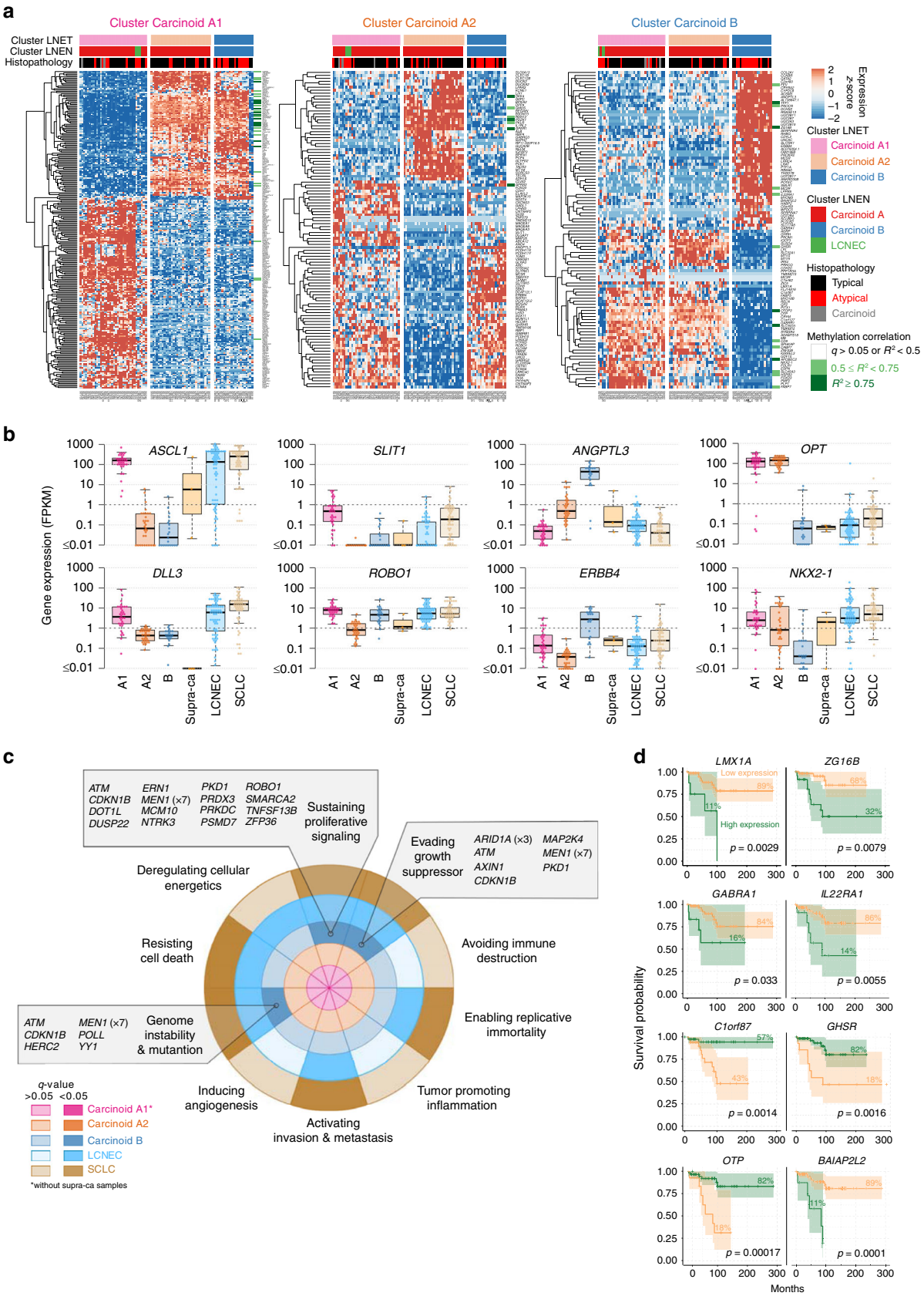

a
Cluster Carcinoid A1
Cluster Carcinoid A2
Cluster Carcinoid B
Cluster LNET
Cluster LNEN
Histopathology
Expression z-score
Cluster LNET
Carcinoid A1
Carcinoid A2
Carcinoid B
Cluster LNEN
Carcinoid A
Carcinoid B
LCNEC
Histopathology
Typical
Atypical
Carcinoid
Methylation correlation
q > 0.05 or R² <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

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

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

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

Discussion

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

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

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

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

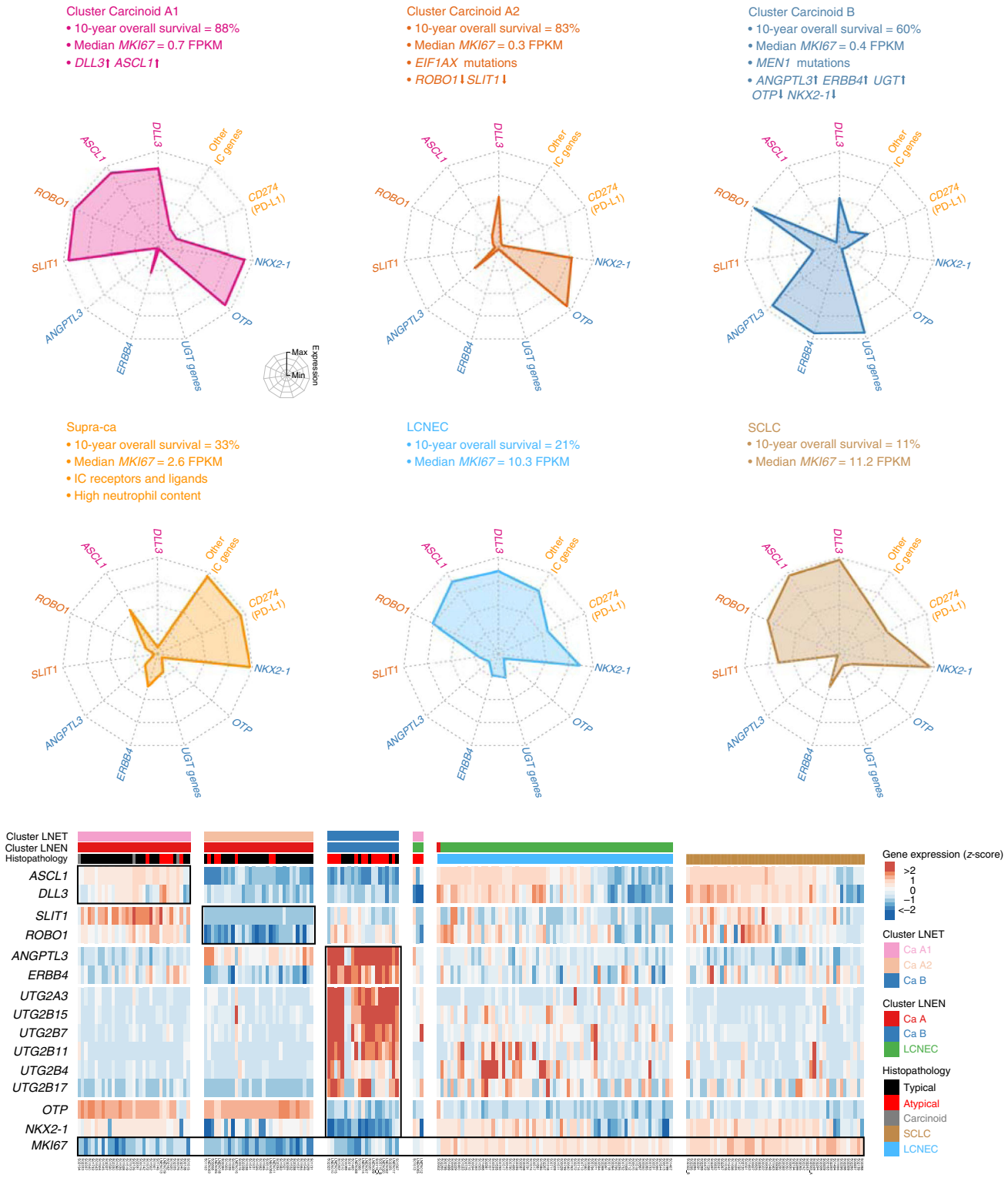


Fig. 6 Main molecular and clinical characteristics of lung neuroendocrine neoplasms. Upper panel: Radar charts of the expression level (*z*-score) of the characteristic genes [*DLL3, ASCL1, ROBO1, SLIT1, ANGPTL3, ERBB4*, UGT genes family, *OTP, NKX2-1, PD-L1 (CD274)*, and other immune checkpoint genes]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

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

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

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

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

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

Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In order to test whether using additional latent factors could increase the power to detect molecular clusters, we performed a similar analysis but using all five latent factors identified by MOFA. In order to provide more importance to the factors most likely to capture the biological variation in the data, the multiple replicate clusterings were performed using a weighted k-means algorithm, where variables (here MOFA latent factors) are given weights corresponding to their proportion of variance explained. More specifically, instead of minimising the within-cluster sum of squares, the weighted within-cluster sum of squares is minimised. Results for $K=3$ clusters of LNET and LNEN samples are presented in Supplementary Fig. 8. We can see that the alternative approach (weighted K-means on five latent factors) leads to the exact same cluster membership as the original approach (K-means on $K-1$ latent factors), both for LNEN and LNET clusters. Indeed, among the latent factors, only the first 3 were associated with either the LNEN clusters (ANOVA $q=4.09\times10^{-84}$, 8.63×10^{-80}, 0.66, 0.094, 0.24, respectively, for latent factors 1 through 5) or the LNET clusters (ANOVA $q=5.06\times10^{-4}$, 5.99×10^{-47}, 5.12×10^{-46}, 0.15, 0.052, respectively), which indicates that the first three latent factors captured the differences between clusters. The code used for the clustering analyses (integration_unsupervised.R) is available at https://github.com/IARCbioinfo/integration_analysis_scripts.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Data availability

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

Code availability

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

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

References

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

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

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

Acknowledgements

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

Author contributions

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

Additional information

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

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

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

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

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

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

ORIGINAL ARTICLE

Gene Expression Profiling of Lung Atypical Carcinoids and Large Cell Neuroendocrine Carcinomas Identifies Three Transcriptomic Subtypes with Specific Genomic Alterations

Michele Simbolo, PhD,[a] Stefano Barbi, PhD,[a] Matteo Fassan, MD,[b] Andrea Mafficini, PhD,[b] Greta Ali, MD,[c] Caterina Vicentini, PhD,[a,b] Nicola Sperandio, DScTech,[a,b] Vincenzo Corbo, PD,[a] Borislav Rusev, MD,[b] Luca Mastracci, MD,[d] Federica Grillo, MD,[d] Sara Pilotto, MD, PhD,[e] Giuseppe Pelosi, MD,[f] Serena Pelliccioni, MD,[c] Rita T. Lawlor, PhD,[b] Giampaolo Tortora, MD, PhD,[e,g,h] Gabriella Fontanini, MD,[c] Marco Volante, MD,[i] Aldo Scarpa, MD, PhD,[a,b,*] Emilio Bria, MD[g,h]

[a]*Department of Diagnostics and Public Health, Section of Anatomical Pathology, University and Hospital Trust of Verona, Verona, Italy*
[b]*ARC-Net Research Centre, University and Hospital Trust of Verona, Verona, Italy*
[c]*Department of Surgical, Medical, Molecular Pathology and Critical Area, University of Pisa, AOU Pisana, Pisa, Italy*
[d]*Department of Surgical and Diagnostic Sciences, University of Genoa and IRCCS S. Martino-IST University Hospital, Genoa, Italy*
[e]*Department of Medicine, Section of Medical Oncology, University and Hospital Trust of Verona, Verona, Italy*
[f]*Department of Oncology and Hemato-Oncology, University of Milan, and Inter-Hospital Pathology Division, IRCCS MultiMedica, Milan, Italy*
[g]*Comprehensive Cancer Center, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy*
[h]*Sacred Hearth Catholic University, Rome, Italy*
[i]*Department of Oncology, University of Turin at San Luigi Hospital, Orbassano, Turin, Italy*

Received 11 January 2019; revised 25 March 2019; accepted 6 May 2019
Available online - 11 May 2019

ABSTRACT

Introduction: DNA mutational profiling showed that atypical carcinoids (ACs) share alterations with large cell neuroendocrine carcinomas (LCNECs). Transcriptomic studies suggested that LCNECs are composed of two subtypes, one of which shares molecular anomalies with SCLC. The missing piece of information is the transcriptomic relationship between ACs and LCNECs, as a direct comparison is lacking in the literature.

Methods: Transcriptomic and genomic alterations were investigated by next-generation sequencing in a discovery set of 14 ACs and 14 LCNECs and validated on 21 ACs and 18 LCNECs by using custom gene panels and immunohistochemistry for Men1 and Rb1.

Results: A 58-gene signature distinguished three transcriptional clusters. Cluster 1 comprised 20 LCNECs and one AC harboring concurrent inactivation of tumor protein p53 gene (*TP53*) and retinoblastoma 1 gene (*RB1*) in the absence of menin 1 gene (*MEN1*) mutations; all cases lacked Rb1 nuclear immunostaining. Cluster 3 included 20

***Corresponding author.**

Disclosure: Dr. Bria has received honoraria or speakers' fees from MSD, AstraZeneca, Celgene, Pfizer, Helsinn, Eli Lilly, BMS, Novartis, and Roche, as well as research grants from Italian Association for Cancer Research (Associazione Italiana Ricerca sul Cancro, AIRC), the International Association for the Study of Lung Cancer (IASLC), Lega Italiana per la Lotta contro i Tumori (LILT), Cariverona Foundation (Fondazione Cariverona), AstraZeneca, Roche, and Open Innovation outside the submitted work. Dr. Pilotto reports personal fees from AstraZeneca, Eli Lilly, BMS, Boehringer Ingelheim, Roche, MSD and Istituto Gentili outside the submitted work. Dr. Tortora reports grants and personal fees from Celgene, Novartis, Roche, Incyte, and Merck Serono for serving on advisory boards and as a consultant, as well as grants from AIRC and Fondazione Cariverona outside the submitted work. The remaining authors declare no conflict of interest.

Address for correspondence: Aldo Scarpa, MD, PhD, University of Verona, Polyclinic Gian Battista Rossi, Piazzale Ludovico Antonio Scuro 10, Verona 37134, Italy. E-mail: aldo.scarpa@univr.it

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2019.05.003

ACs and four LCNECs lacking *RB1* alterations and having frequent *MEN1* (37.5%) and *TP53* mutations (16.7%); menin nuclear immunostaining was lost in 75% of cases. Cluster 2 included 14 ACs and eight LCNECs showing intermediate features: *TP53*, 40.9%; *MEN1*, 22.7%; and *RB1*, 18.2%. Patients in cluster C1 had a shorter cancer-specific survival than did patients in C2 or C3.

Conclusions: ACs and LCNECs comprise three different and clinically relevant molecular diseases, one AC-enriched group in which *MEN1* inactivation plays a major role, one LCNEC-enriched group whose hallmark is *RB1* inactivation, and one mixed group with intermediate molecular features. These data support a progression of malignancy that may be traced by using combined molecular and immunohistochemical analysis.

Keywords: Lung neuroendocrine tumors; Atypical carcinoid; Large cell neuroendocrine carcinoma; Gene expression profiling; Next-generation sequencing; Transcriptomics

Introduction

The current WHO classification divides lung neuroendocrine tumors (LNETs) into four histological variants: typical carcinoid (TC), atypical carcinoid (AC), large cell neuroendocrine carcinoma (LCNEC), and SCLC.[1] From a clinical standpoint, carcinoids (TCs and ACs) are distinguished from carcinomas (LCNEC and SCLC). TCs are low-grade tumors, with patients having a long life expectancy, and ACs are intermediate-grade tumors with varying clinical behavior. Conversely, LCNECs and SCLCs are both high-grade tumors with a dismal prognosis.[1–3]

The insufficient knowledge of LNET biology limits the comprehension of these tumor subtypes, which to date have been considered as having a separate pathogenesis.[4–7] Recent DNA mutational profiles showed that carcinoids and carcinomas share similar gene alterations, but with different prevalence among the subtypes.[8] Alterations in chromatin remodeling genes are found in all four variants, whereas menin 1 gene (*MEN1*) alterations are found mainly in carcinoids, and inactivation of tumor protein p53 gene (*TP53*) and retinoblastoma 1 gene (*RB1*) is significantly enriched in carcinomas.[5,8–12] The fact that the same gene alterations found in carcinomas are identified in low-grade tumors but at a lower prevalence may suggest the existence of progression of malignancy and the development of secondary high-grade neuroendocrine carcinomas from preexisting carcinoids.[8,13]

Gene expression profiles produce a global picture of cellular function, and it has been shown that the transcriptional phenotypes of lung cancers mimic the WHO classification.[6,14] They may also provide additional stratification within histological subtypes (that is, the potential to identify molecular subgroups within tumors showing similar morphological features).[14,15]

Three gene expression profiling studies of LNETs have been recently published.[9,14,16] Asiedu et al. reported that transcriptomic profiles could distinguish between carcinoids (TCs and ACs) and SCLC[16]; in this study LCNECs were not included. Karlsson et al. analyzed an ample series of large cell lung carcinomas, including LNECs, and observed a clear separation of three transcriptional groups: adenocarcinoma, squamous cell carcinoma, and a third neuroendocrine group comprising SCLC and LCNEC.[14] Moreover, the comparison of LCNEC and SCLC showed that LCNEC exhibited two different transcriptional profiles associated with different *TP53* and *RB1* genes alteration patterns, corresponding to a proposed genetic division of LCNEC into SCLC-like and NSCLC-like cancers.[14] A study of 75 LCNECs by George et al. confirmed the existence of two LCNEC subtypes, one (type II) characterized by the concurrent inactivation of *TP53* and *RB1* (42%) and one (type I) with *TP53* and serine/threonine kinase 11 gene (*STK11*)/kelch like ECH associated protein 1 gene (*KEAP1*) alterations (37%), but clearly showed that LCNECs have no transcriptional relationship with adenocarcinomas and squamous cell carcinomas.[9]

The missing piece of information is the transcriptomic relationship between AC and LCNEC, as a direct comparison, is lacking in the literature.

Materials and Methods

Cases

A cohort of 67 surgically resected LNETs was collected from four Italian institutions (Applied Research on Cancer-Network [ARC-Net] Research Centre-Verona, IRCCS San Martino-Genova, University of Pisa, and AUO Orbassano-University of Turin). All cases were reclassified according to the WHO 2015 criteria[1] and included 35 ACs and 32 LCNECs. Neuroendocrine differentiation was assessed by using immunostaining for chromogranin, synaptophysin, and CD56.[1,17,18] The AC diagnostic criteria included a well-differentiated morphology with between two and 10 mitoses per 2 mm^2 of area and/or presence of focal necrosis.[19,20] LCNECs were diagnosed on the basis of non–small cell cytologic features, including large cell size, low nuclear-to-cytoplasmic ratio, prominent nucleoli or vesicular chromatin, a mitotic rate of more than 10 mitoses per 2 mm^2 (average 60–80 mitoses per 2 mm^2), and more extensive necrosis.[18–20] Tumor stage was assigned according to the seventh edition of the TNM classification

of malignant tumors.[21] None of the patients received preoperative therapy. Samples were divided into two groups: a discovery set including 28 samples (14 ACs and 14 LCNECs) and a validation set of 39 samples (21 ACs and 18 LCNECs).

Ethics

Ethics committee approval (ECA) was obtained at the four institutions: ARC-Net Research Centre-Verona (ECA no. 2173-prot.26775 [1 June 2012]), AUO Orbassano-University of Torino (ECA no. 167/2015-prot. 17975 [October 21, 2015]), IRCCS San Martino-Genova (ECA no. 027/2016LM [16 March 2016]), and University of Pisa [ECA no. 1040/16 [March 31, 2016]).

Mutational, CNV, and Expression Analysis by Next-Generation Sequencing

The details of the experimental procedures are described in the Supplementary Methods. Briefly, nucleic acids were obtained from formalin-fixed paraffin-embedded tissues as reported.[22,23] Sequencing was performed on Ion Torrent platform (Thermo Fisher Scientific). Data analysis including variant calling was done by using Torrent Suite Software, version 5.0 (Termo Fisher Scientific). Unfiltered variants are reported in Supplementary Tables 1 and 2. Filtered variants were annotated by using a custom pipeline based on vcflib (https://github.com/ekg/vcflib), SnpSift,[24] and Variant Effect Predictor.[25] Annotated variants were filtered by using only the canonical transcripts. Only missense, nonsense, frameshift, or splice site variants were retained. Germline variants were removed. Alignments were visually verified with the Integrative Genomics Viewer, version 2.3,[26] to confirm the presence of identified mutations and to exclude sequencing artefacts. The mutational profile[27] of each sample was obtained with the MuSiCa software.[28] Copy number variation (CNV) was evaluated by using OncoCNV software, version 6.8.[29] The AmpliSeqRNA plugin was used to analyze expression profiling data. Differential analysis was performed by using the DESeq2[30] package for R; an adjusted p value less than 0.05 was considered significant. For gene set analysis, the GSVA[31] package was used.

Immunohistochemistry

Immunostaining was performed by using the Bond Polymer Refine Detection kit (Leica Biosystems) in a BOND-MAX system (Leica Biosystems) on 4-μm-thick formalin-fixed paraffin-embedded sections with the primary antibodies for menin (clone A300-105A [Bethyl Laboratories], dilution 1:1000) and Rb (clone 4H1 [Cell Signaling Technology], dilution 1:250). Appropriate positive and negative controls were run concurrently.

Statistical Analysis

One-way analysis of variance, the Kruskal-Wallis test, the Fisher test with Monte Carlo simulation, and the Fisher exact test were used as appropriate; correction for multiple comparisons was performed according to Benjamini-Hochberg. The Mantel-Cox test was used to compare survival curves. A p value less than 0.05 was considered as significant. Analyses were performed by using Medcalc for Windows, version.18.11 (MedCalc Software), and R software, version 3.5.3.[32]

Results

Study Design

The cohort of 67 cases was divided into a discovery set and a validation set, consisting of 28 and 39 cases, respectively. The study workflow is depicted in Supplementary Figure 1.

The discovery screen was performed on 14 ACs and 14 LCNECs with use of the Ampliseq Transcriptome Human Gene Expression Kit (ThermoFisher), which investigates the expression of 20,815 human genes, and the Ampliseq Comprehensive Cancer Panel (ThermoFisher) for mutational and CNV analysis of 409 genes.

The validation set comprised 21 ACs and 18 LCNECs and was analyzed by targeted sequencing with the use of two custom panels. The first panel was designed to assess the mRNA expression level of 60 genes, including 58 that were differentially expressed in the discovery set plus *MEN1* and *RB1*. The second panel was devised to evaluate DNA alterations in 16 genes: seven genes for mutational analysis only, three genes for CNV analysis only, and six genes for both mutational and CNV analysis.

The expression profiles clustering analysis and the prevalence of mutations and CNVs were finally computed on the entire cohort of 67 cases.

Clinicopathologic Features

Clinicopathologic data are summarized in Supplementary Table 3 and detailed in Supplementary Table 4. The 67 patients had a mean age of 66.2 years and a median clinical follow-up time of 17 months (range 2–100). Of the 67 cases, 33 (49%) were stage I, 24 (36%) stage II, six (9%) were stage III, and four (6%) were stage IV. ACs and LCNECs differed by patient age, patient sex, tumor size, and Ki67 index, whereas there was no statistically significant difference for smoking status and stage (see Supplementary Table 3).

Gene Expression Analysis and Unsupervised Hierarchical Clustering of Discovery Set

Gene expression analysis was performed on the 28 samples of the discovery set and eight nonneoplastic lung samples. Hierarchical unsupervised clustering analysis

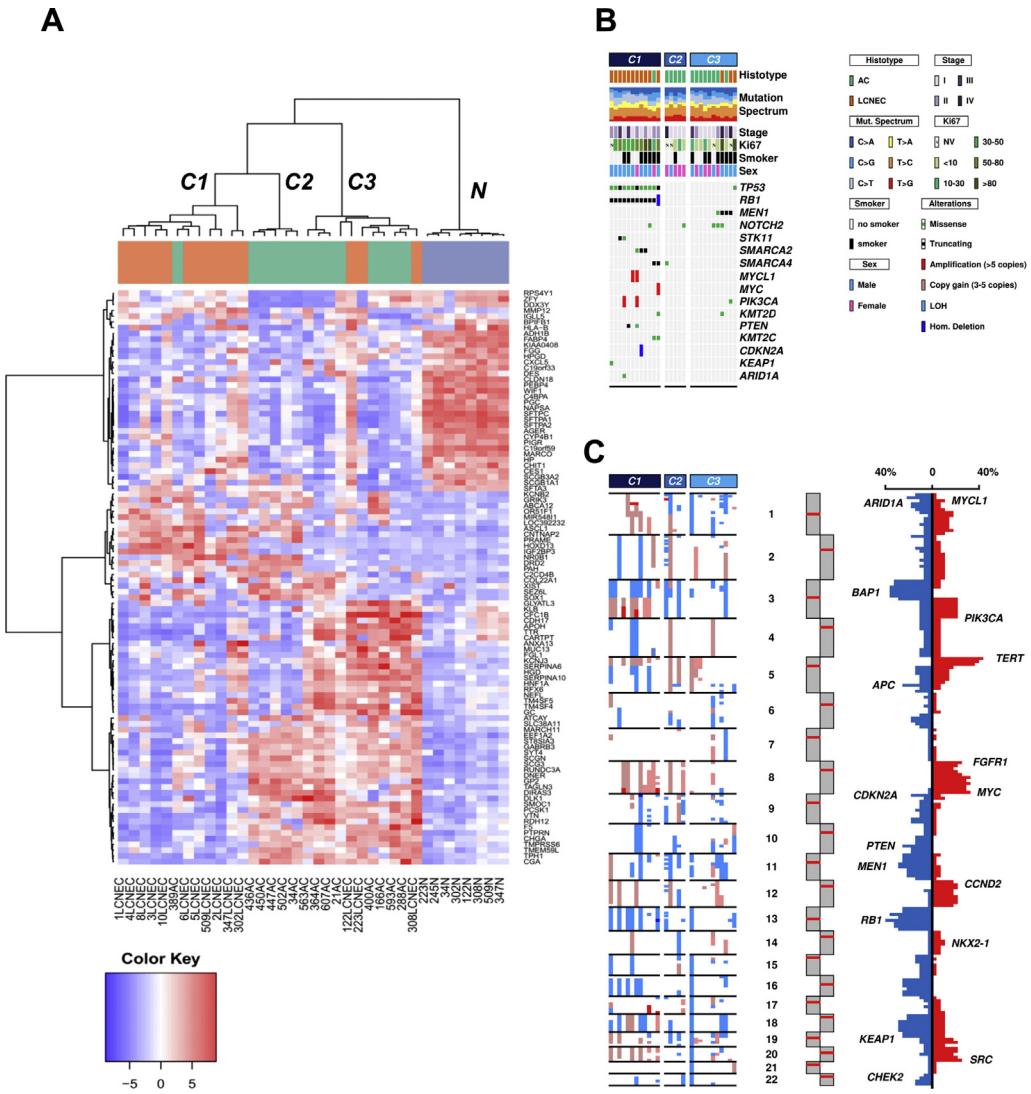


Figure 1. Transcriptome analysis of the discovery set of 28 lung neuroendocrine neoplasms distinguished three molecular clusters of tumors. (*A*) Hierarchical unsupervised clustering of transcriptomes of 14 atypical carcinoids (ACs) (*green*) and 14 large cell neuroendocrine carcinomas (LCNEC) (*orange*) plus eight normal lung (N) (*light blue*) samples with use of the Ward D2 algorithm Tumors were grouped in three separate clusters (C1, C2, and C3) that differ from normal lung samples. Case ID is indicated at the bottom, gene names are indicated on the right. (*B*) Alterations in 16 genes at sequencing analysis; the legend for pathological and molecular alterations is reported in the panel on the right. The mutation spectrum takes into consideration all nonsynonymous variants detected per megabase of exonic sequence, grouped into six classes; stacked bars represent the percentage of each group in each sample. (*C*) Copy number variations within the three clusters (left panel) refer to chromosomes; frequency of copy number variation alterations (*right panel*) where copy gain events are indicated in red and losses in blue. *TP53*, tumor protein p53 gene; *RB1*, retinoblastoma 1 gene; *MEN1*, menin 1 gene; *NOTCH2*, notch 2 gene; *STK11*, serine/threonine kinase 11 gene; *SMARCA2*, SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2 gene; *SMARCA4*, SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 gene; *MYCL1*, v-myc avian myelocytomatosis viral oncogene lung carcinoma derived homolog gene; *MYC*, v-myc avian myelocytomatosis viral oncogene homolog gene; *PIK3CA*, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha gene; *KMT2D*, lysine methyltransferase 2D gene; *PTEN*, phosphatase and tensin homolog gene; *KMT2C*, lysine methyltransferase 2C gene; *CDKN2A*, cyclin dependent kinase inhibitor 2A gene; *KEAP1*, kelch like ECH associated protein 1 gene; *ARID1A*, AT-rich interaction domain 1A gene; *BAP1*, BRCA1 associated protein 1 gene; *TERT*, telomerase reverse transcriptase; *APC*, APC, WNT signaling pathway regulator; *FGRF1*, gene; *CCND2*, cyclin D2 gene; *NKX2-1*, NK2 homeobox 2 gene; *SRC*, SRC proto-oncogene, non-receptor tyrosine kinase gene; *CHEK2*, checkpoint kinase 2 gene.

using the Ward D2 algorithm identified four clusters (Fig. 1*A*): cluster 1 (C1), which included 12 samples (11 LCNECs and one AC); cluster 2 (C2), which included five samples (all ACs); cluster 3 (C3) , which included 11 samples (eight ACs and three LCNECs); and a fourth group (N), which included all nonneoplastic lung samples.

To verify the robustness of clusters, nonnegative matrix factorization (NMF) based on the expression

profile of the top varying 5000 genes was performed (Supplementary Fig. 2). NMF supported the presence of three LNET classes and one class of normal samples (Supplementary Fig. 2*A*). Comparison of the four NMF classes versus the four hierarchical clustering clusters showed that 32 of 35 cases (91.4%) were consistently assigned to corresponding groups (Supplementary Fig. 2*B*), confirming the reliability of clusters.

To analyze differentially affected pathways, gene set variation analysis was performed[31] aggregating gene expression data according to the "hallmark" gene sets collection from the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb). Gene set variation analysis identified differentially enriched gene sets between tumor clusters (Supplementary Fig. 3). In particular, the gene sets of E2F targets and G2/M checkpoint showed a progressively higher score moving from cluster C3 toward cluster C1 ($p < 0.001$). The E2F targets gene set includes downstream targets of the E2F transcription factors family, which play a major role in G1/S transition.[33] Similarly, genes of the G2M checkpoint set mediate progression through the cell cycle. Thus, coordinated enrichment of both sets is consistent with increased proliferation.

Conversely, the bile acid metabolism gene set, including members involved in peroxisome organization, was increasingly up-regulated moving from C1 to C3 ($p = 0.0217$). A similar trend was also observed in the other gene sets enriched at a false discovery rate of 0.1. Indeed, genes involved in the mitotic spindle and v-myc avian myelocytomatosis viral oncogene homolog gene (*MYC*) targets were enriched in C1, consistent with recurrent *MYC* copy gain in this cluster, whereas gene sets related to bile acid metabolism (fatty acid metabolism, xenobiotic metabolism, and peroxisome) were enriched in C3. Finally, C1 and C2 displayed enrichment in Wnt signalling compared with C3 (Supplementary Table 5 and see also Supplementary Fig. 3).

A set of 58 genes was identified as differentially expressed among the three clusters ($p < 0.05$). The details on differentially expressed genes are reported in Supplementary Table 6 and their distribution in the three LNET clusters in Supplementary Figure 4.

A 58-Gene Signature Identifies Three Expression Profiling Clusters with Distinct Clinicopathologic Features

An RNA targeted sequencing custom panel, designed by using the 58-gene signature identified in the discovery set, was used to analyze the entire series of 67 cases (35 ACs and 32 LCNECs) comprising the 28 of the discovery set and the 39 of the validation set. Additionally, *MEN1* and *RB1* transcripts were included in the custom panel because of their known involvement in AC and LCNEC, respectively,[9,14,34] to correlate their expression levels with the mutational status.

Hierarchical clustering using the 58 genes and Ward D2 algorithm categorized the cases in three clusters (Fig. 2), which were consistent with those obtained by the analysis of 20,815 genes in the discovery set. Clinicopathologic features of the 67 cases are compared across the three clusters in Table 1. Cluster 1 contained 20 LCNECs and one AC. This cluster showed a higher Ki67 index (mean 66% [$p < 0.0001$]) and shorter cancer-specific survival ($p = 0.26$). C3 included 20 ACs and four LCNECs characterized by a lower Ki67 index (mean 21%) and did not reach the median cancer-specific survival (Supplementary Fig. 5). C2 included 14 ACs and eight LCNECs showing intermediate features between those in C1 and C3, with a mean Ki67 index of 36% and a median cancer-specific survival of 47 months. Remarkably, this intermediate C2 cluster was composed of two subclusters (see Fig. 2). C2a included eight LCNECs and three ACs, the former characterized by *TP53* mutations in all eight cases associated with a heterozygous *RB1* mutations in four of them, whereas one of the three ACs had a *MEN1* mutation. C2b included 11 ACs, four of which harbored a *MEN1* mutation and one had a *TP53* mutation. The Ki67 index in C2a (mean 60.0, median 60.0, range 10%–80%) was higher than in C2b (mean 12.0; median 7.0; range 3-35%).

MEN1 and *RB1* expression levels were differentially distributed among the clusters, with significant underexpression of *RB1* in all C1 cases and significant underexpression of *MEN1* in most of the C3 samples and one-third of the C2 samples (Fig. 3*A*).

Discovery Screen of Mutations and CNVs of 14 ACs and 14 LCNECs

Mutational analysis was performed on the discovery set for the coding sequence of 409 genes. Sequencing achieved an average coverage of 698× (198×–1657×) in tumor and 386× (26×–981×) in normal samples (Supplementary Table 7).

Mutations were identified in 22 of the 28 cases. All 12 cases in C1 harbored mutations, whereas two of five cases in C2 and four of 11 in C3 had no mutations. A total of 79 mutations were identified in 36 genes: 56 missense, 10 nonsense, eight frameshift, and five splice site (Supplementary Table 8). *TP53* was the most frequently mutated gene (13 of 28 [46.4%]), followed by *RB1* (11 of 28 [39.3%]), notch 2 gene (*NOTCH2*) (five of 28 [17.9%]), and *MEN1* (four of 28 [14.3%]) (Fig. 1*B*). The mutational spectrum was prevalently characterized by T>C and C>T transitions, with different relative contributions in individual tumors.

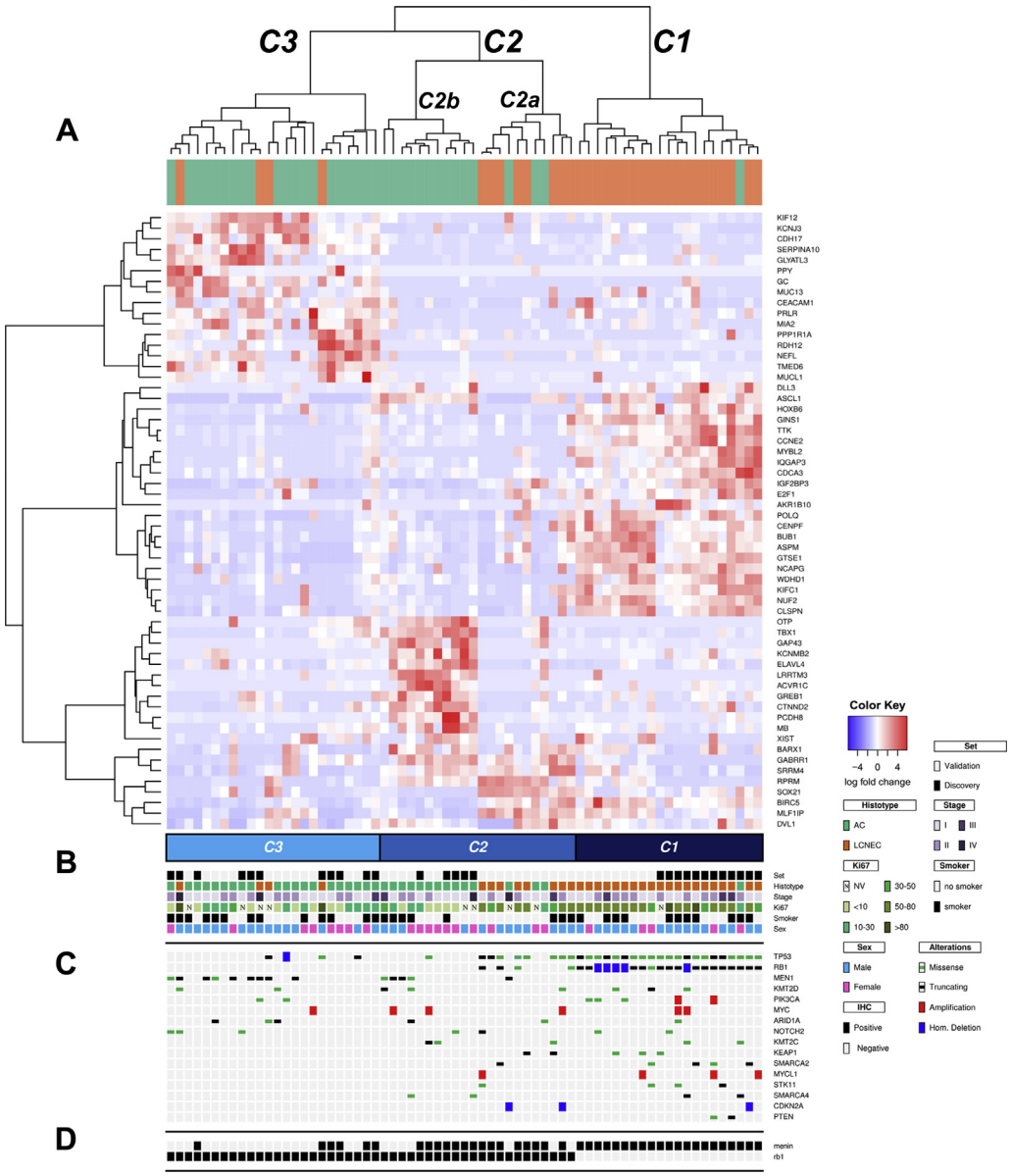


Figure 2. Validation analysis on 67 lung neuroendocrine neoplasms confirms the existence of three clusters of tumors. (*A*) Hierarchical clustering of 35 atypical carcinoids (ACs) and 32 large cell neuroendocrine carcinomas (LCNECs) with use of an RNA custom panel of 58 genes confirmed the presence of the three different molecular subgroups identified by whole transcriptome analysis (see Fig. 1 and Supplementary Fig. 4) and suggests further splitting of cluster 2. (*B*) Clinicopathologic features of the 67 samples; the legend for clinical pathological and molecular alterations is reported in the panel on the right. (*C*) The 16 genes that were altered at sequencing analysis; the legend for alteration type is reported in the panel on the right. (*D*) Immunohistochemical analysis data of menin and Rb.

The most frequent trinucleotide substitution was C[T>C]G, followed by A[T>C]G and A[C>T]G. In particular, all three clusters showed C[T>C]G as the most frequent substitution. The A[T>C]G substitution was the second most frequent in C1 and C2 but not in C3, in which the second most frequent substitution was A[C>T]G. However, no specific substitution was predominant in any cluster.

The CNV status was estimated for all 409 genes by using sequencing data (Supplementary Table 9). The most frequently altered were 20 genes (Supplementary Table 10), including gains in succinate dehydrogenase complex flavoprotein subunit A gene (*SDHA*), RPTOR independent companion of MTOR complex 2 gene (*RICTOR*), telomerase reverse transcriptase gene (*TERT*) (12 of 28 each [42.9%]), and *MYC* (11 of 28 [39.3%]) and losses in BRCA1 associated protein 1 gene (*BAP1*) (12 of 28 [42.9%]), *RB1* (10 of 28 [35.7%]), and *MEN1* (eight of 28 [28.6%]). On the basis of the chromosomal position of each gene, the status of chromosome arms was inferred

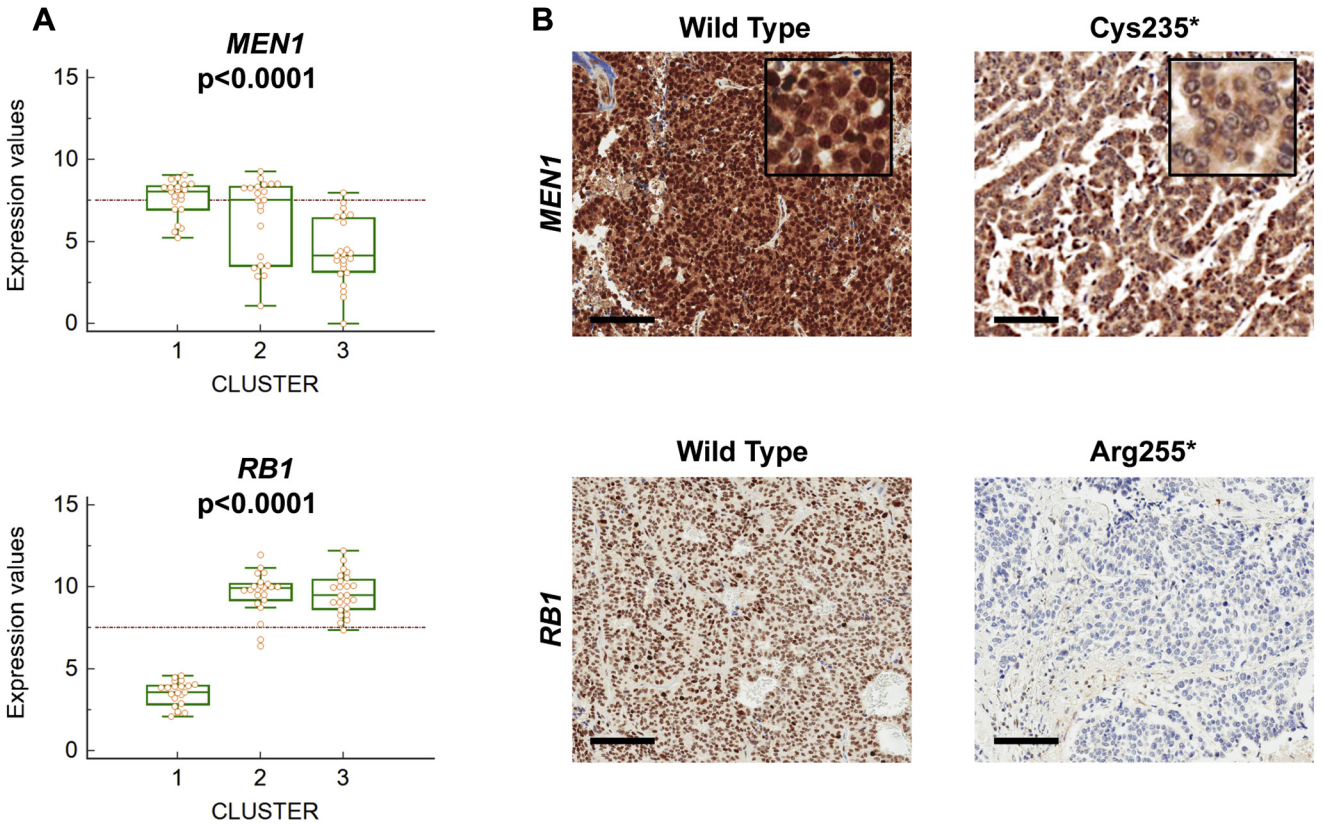


Figure 3. mRNA and immunohistochemical expression analysis of *MEN1* and *RB1*. (*A*) Expression levels (natural logarithm values on *y* axes) of menin 1 gene (*MEN1*) and retinoblastoma 1 gene (*RB1*) genes in the three molecular clusters (on *x* axes) identified by gene expression profiling (see Fig. 2); the dashed red line represents the average gene expression across the entire cohort (reference line). (*B*) Representative images of positive and negative nuclear immunostaining for menin and Rb proteins in two cases harboring a truncating mutation of the corresponding gene, namely Cys235* truncating mutation in *MEN1* gene and Arg255* truncating mutation in *RB1* gene. (scale bars = 100 μm; original magnifications, $\times$10 and $\times$40 [inset]).

(Fig. 1*C* and Supplementary Table 11). Such analysis showed major alterations, namely, losses of chromosome arm 3p (14 of 28 [50.0%]) and whole chromosomes 18 (11 of 28 [39.3%]) and 11 (nine of 28 [32.1%]) and gains in chromosome arms 8q (12 of 28 [42.9%]) and 5p (11 of 28 [39.3%]). Chromosome alterations according to tumor type and expression profile clusters are illustrated in Supplementary Figure 6.

Validation of Mutations and CNVs in 21 ACs and 18 LCNECs by Targeted Sequencing of 16 Genes

The validation set was analyzed by targeted sequencing using a DNA custom panel of 16 genes altered in both the present study and other studies investigating LNETs,[5,9,16,35] investigating the mutational status of seven genes, CNV of three genes, and both alterations of six genes (Supplementary Table 12). Sequencing achieved an average coverage of 1649$\times$ (329$\times$–3636$\times$) in tumor and 468$\times$ (165$\times$–1643$\times$) in normal samples (see Supplementary Table 7). A total of 67 mutations were identified: 38 missense mutations, 13 nonsense mutations, 11 frameshift mutations, three in-frame deletions, and two splice site mutations (see Supplementary Table 8). *TP53* was the most frequently mutated gene (20 of 39 [51.3%]), followed by *MEN1* (10 of 39 [25.6%]) and *RB1* (nine of 39 [23.1%]).

Analysis of CNV status for the nine selected genes identified gains in *MYC* (five of 39 [12.8%]) and losses in *RB1* (four of 39 [10.3%]) as the most frequent alterations. The comparison between discovery and validation set is reported in Supplementary Table 13.

Prevalence of Gene Mutations and Copy Number Alterations in the Three Different Molecular Clusters

Considering the whole series of 67 cases (35 ACs and 32 LCNECs), the three expression profile clusters showed differences in mutation and CNV frequency for specific genes (Fig. 2*C*). Mutations were present in 56 of 67 cases (see Supplementary Table 8). All 21 patients in C1 harbored mutations, whereas three of 22 in C2 and eight of 24 in C3 had no mutations. Different distributions were identified for *TP53* and *RB1* alterations (each $p < 0.0001$), SWI/SNF related, matrix associated, actin

dependent regulator of chromatin, subfamily a, member 2 gene (*SMARCA2*) ($p = 0.043$) and *MEN1* ($p = 0.008$) mutations, and phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha gene (*PIK3CA*) ($p = 0.035$) alterations (Table 2). C1 was characterized by concurrent *TP53* and *RB1* alterations (21 of 21; 100%); also peculiar to this cluster were alterations in *PIK3CA* (five of 21 [23.8%], $p = 0.038$) and *SMARCA2* (four of 21 [19.0%], $p = 0.047$). C2 showed *TP53* alterations as the most frequent event (nine of 22 [40.9%]), followed by *MEN1* (five of 22 [22.7%]) and *RB1* (four of 22 [18.2%]). C3 had *MEN1* mutation as the most frequent event (nine of 24 [37.5%]), followed by *TP53* alterations (four of 24 [16.7%]), whereas no *RB1* anomaly was displayed.

Expression Levels and Immunohistochemistry of Menin and Rb

Menin and Rb mRNA levels were assessed in all 67 samples. For each sample, normalized log-transformed next-generation sequencing counts were compared between mRNA expression clusters by Kruskal-Wallis test. *RB1* and *MEN1* mRNAs showed differential expression between clusters (each $p < 0.0001$) (see Fig. 3*A*). In particular, *MEN1* mRNA level was very low in 24 samples, of which 17 were in C3 (17 of 24; 70.8%) and seven were in C2 (seven of 22; 31.8%), whereas *RB1* mRNA level was very low in all C1 samples (see Fig. 3*A*). All 14 samples harboring *MEN1* mutations showed very low *MEN1* mRNA.

All cases were immunostained for menin and Rb, and nuclear negativity was interpreted as abnormal (representative cases in Fig. 3*B*). Both *MEN1* and *RB1* displayed strong correlation between mRNA levels and protein immunolabeling ($p < 0.00001$ [Supplementary Fig. 7]).

Lack of menin nuclear immunostaining was detected in 24 of 67 cases (35.8%), including the seven samples in C2 and the 17 in C3 that had low mRNA levels, whereas all cases in C1 had positive immunostaining. A direct correlation between presence of mutations, low mRNA level, and loss of protein nuclear immunostaining was observed ($p < 0.0001$). In detail, all 14 cases with *MEN1* mutation, including 11 ACs and three LCNECs, had low mRNA levels and negative nuclear immunostaining; of these 14 cases, the three LCNECs and six ACs were in C3 whereas five ACs were in C2. Interestingly, 10 cases were wild-type for *MEN1* at sequencing; nine ACs belonging to C3 and one LCNEC belonging to C2, showed loss of nuclear menin, suggesting the existence of additional mechanisms of *MEN1* inactivation.[34] All cases in C1 were *MEN1* wild-type at sequencing and had positive menin immunostaining.[36]

Lack of Rb nuclear immunostaining was found in all 21 samples of C1 (20 LCNECs and one AC), with concomitant low mRNA levels and biallelic inactivation of *RB1* owing to homozygous deletion in five cases or heterozygous mutation and loss of the wild-type allele in 16 cases (Supplementary Table 14 and see also Fig. 2). Conversely, all samples in C2 and C3 had positive nuclear immunostaining, including four cases in C2 with *RB1* heterozygous mutations and retention of the wild-type allele.

Discussion

It has been suggested that high-grade neuroendocrine carcinomas may represent a progression of malignancy of preexisting carcinoids.[13] Indeed, recent genomic and transcriptomic data indicate that ACs are hybrid tumors sharing genomic features with both low-grade (TCs) and high-grade (LCNECs and SCLCs) neuroendocrine neoplasms,[8] and that LCNECs may be subdivided in at least two molecular subgroups, one of which shows molecular similarities with SCLCs.[7,9,14] This prompted us to perform a direct comparison of molecular alterations of ACs and LCNECs, which is lacking in literature.

The comparative transcriptomic analysis of ACs and LCNECs reported herein discriminated three transcriptional clusters, defined as C1, C2, and C3, which also showed specific genomic patterns (Fig. 4).

C3 was an AC-enriched cluster that included 20 ACs (83.3%) and four LCNECs (16.7%). *MEN1* mutations were the most frequent events (nine of 24 [37.5%]), followed by *TP53* mutations (16.7%). No case had *RB1* alterations. Interestingly, three of the four LCNECs in this cluster harbored *MEN1* mutations, which are relatively rare in LCNECs, in which they account for 4% of cases.[8,35] That *MEN1* alterations may represent a major event in this cluster is suggested by loss of menin nuclear immunostaining in most cases (18 of 24 [75.0%]), including nine samples harboring a *MEN1* mutation and nine that were determined to be wild-type, suggesting the existence of additional inactivation mechanisms.[34] Indeed, the immunohistochemical findings in LNETs of the present series parallel those in pancreatic neuroendocrine neoplasms, in which of 80% of cases lacking menin nuclear immunostaining, only 30% revealed *MEN1* mutations by sequencing analysis,[36] and subsequent whole-genome analysis showed gross chromosome 11 alterations in many cases.[37]

C1 was an LCNEC-enriched cluster consisting of 20 LCNECs and one AC. All cases in this cluster had concurrent inactivation of *TP53* and *RB1* genes (100%) and lacked *MEN1* mutations. All samples showed low *RB1* mRNA and loss of nuclear immunostaining for Rb protein. Other frequent alterations in C1 were found in

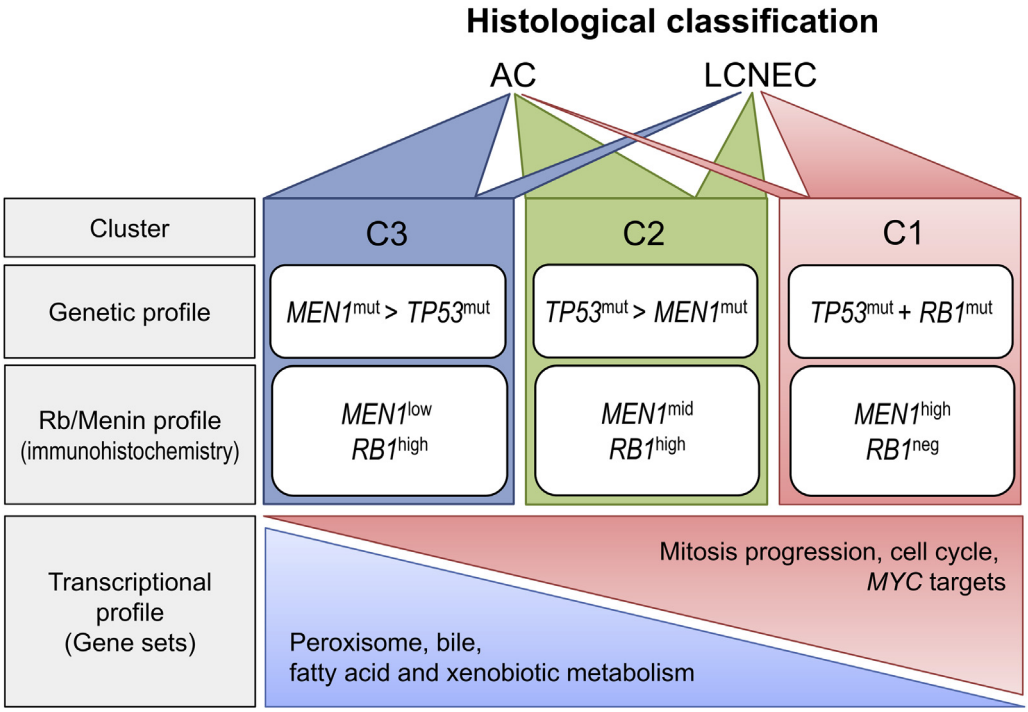


Figure 4. Outline of main differences between the three clusters of the lung neuroendocrine tumors. Cluster 3 is mainly composed of atypical carcinoids (ACs), is Rb proficient, features frequent menin 1 gene (*MEN1*) and rare retinoblastoma 1 gene (*TP53*) mutations, and displays high levels of oxidative metabolism. Cluster 1 is composed almost exclusively of large cell neuroendocrine carcinomas (LCNECs), has Rb loss of expression, always features *TP53* and *RB1* inactivation, and displays high levels of cell cycle deregulation. Cluster 2 has an intermediate transcriptional profile compared with C1 and C3 and is composed of both ACs (one-third of which bear a *MEN1* mutation) and LCNECs (featuring *TP53* mutation but retaining Rb expression). profile: neg, negative staining in all cases; low, negative staining in most cases; mid, similar proportion of positive and negative cases; high, positive staining in most cases; mut, mutation; *MYC*, v-myc avian myelocytomatosis viral oncogene homolog gene.

SMARCA2 (19.0%), *STK11, KEAP1,* and v-myc avian myelocytomatosis viral oncogene lung carcinoma derived homolog gene (*MYCL1*) (each 14.3%) genes.

C2 showed intermediate features between C1 and C3. This group included 14 ACs (64%) and eight LCNECs (36%), in which mutations in *TP53* was the most frequent event (41%), followed by mutations in *MEN1* (23%) and *RB1* (18%) genes. Menin nuclear immunostaining was lost in six samples, comprising the five harboring *MEN1* mutations and one *MEN1* wild-type. Interestingly, nuclear immunostaining of Rb was retained in all cases in this cluster, including the four that harbored a mutated allele but retained the normal allele, supporting the central role of the biallelic inactivation of *RB1* in the transcriptomic shift from C2 to C1.

A comparison of our clusters with the subsets of Rekhtman et al.,[35] and George et al.,[9] both comprising only LCNECs, has identified some interesting overlap. Of note, Rekhtman et al. defined those LCNECs harboring *MEN1* mutations as a third minor subset of "carcinoid-like" LCNECs, and these cases were part of our AC-predominant C3.[35] Our C1 LCNEC-enriched cluster with concurrent inactivation of *TP53* and *RB1* genes and lacking *MEN1* mutations coincide with both the SCLC-like LCNEC subtype of Rekhtman et al.[35] and the type II LCNECs of George et al.[9] Interestingly, other frequent alterations found in our C1 were *SMARCA2*, *STK11, KEAP1,* and *MYCL1*. As these genes are frequently altered in NSCLCs, Rekhtman et al. classified cases harboring these alterations and lacking concurrent *RB1/TP53* inactivation as NSCLC-like LCNEC.[35] Similarly, George et al. suggested that these alterations are typical of their type II LCNECs, but clearly reported that they had no transcriptional relationship with NSCLCs.[9] Furthermore, inactivation of *TP53* with retained Rb immunostaining notable to our C2, are characteristics of the NSCLC-like subtype of Rekhtman et al. and the type I of George et al.[9,35] Notably, our finding that these alterations may occur in association with concurrent *RB1/TP53* inactivation is not surprising, as several cases in the series of both Rekhtman et al. and George et al. displayed the same phenomenon.[9,35]

The identification of three clusters of ACs and LCNECs, two of which were enriched for either AC or LCNEC and the third with intermediate features, suggests the existence of a progression of malignancy for a

proportion of ACs into LCNECs. This is further supported by the fact that the intermediate C2 cluster was composed of two subclusters. The first (C2a) was closer to C1 because of both a similarly high Ki67 index (mean 60%) and the presence of LCNECs with *TP53* mutations associated with heterozygous RB1 alterations. The second (C2b), which is closer to C3, is composed entirely of ACs with a mean Ki67 index of 12% and enriched for *MEN1* mutations. The ACs in C2, especially those sharing the C2a subcluster with LCNECs, might overlap with the recently proposed "supracarcinoids," which were identified by Alcala et al. through supervised machine learning of genomic and transcriptomic data.[38] Indeed, supracarcinoids were defined as a subgroup of LNETs with a clear carcinoid histopathologic pattern but with molecular characteristics similar to those of LCNECs.[38] This further supports the hypothesis that carcinoids may evolve into carcinomas by accumulation of genetic anomalies. A recent publication explicitly suggested the existence of an evolution from AC to LCNEC on the basis of clustering of mutations and CNVs.[13] This might be especially true for a fraction of ACs that do not display *MEN1* loss but show *TP53* alterations.

The three clusters described here also differed from a clinicopathologic point of view, including Ki67 proliferation index and cancer-specific survival. C1 had a mean Ki67 index of 66% versus 37% and 21% for C2 and C3, respectively. Follow-up data were available for 56 patients. Patients in C1 had the shortest survival (median 19 months), whereas patients in C2 had a median survival of 47 months and patients in C3 did not reach the median survival during follow-up.

In conclusion, our study shows that ACs and LCNECs comprise three different molecular diseases of potential clinical relevance, one AC-enriched group in which *MEN1* inactivation plays a major role, one LCNEC-enriched group whose hallmark is *RB1* inactivation, and one group with intermediate features. Indeed, it has been reported that carcinoids harboring *MEN1* mutations, loss of heterozygosity, and low mRNA levels had shorter overall survival,[34] whereas the independent poor prognostic role of *RB1* inactivation is common to AC, LCNEC, and SCLC.[8] Molecular profiling with a combined immunohistochemical and mutational analysis using routinely available paraffin-embedded tissues may complement histological examination to provide better diagnostic definition and prognostic stratification of LNETs that would be helpful for their clinical management.

Acknowledgments

This study was funded by the following organizations: Italian Cancer Genome Project, Ministry of University and Research (grant FIRB RBAP10AHJB); Italian Association for Cancer Research (AIRC grants 5x1000 12182 to Dr. Scarpa and Dr. Tortora, IG 19238 to Dr. Volante, and IG 20583 to Dr. Bria); Cariverona Foundation (grant 2015.0872 to Dr. Tortora and Project Antonio Schiavi funding to Dr. Lawlor); and International Association for the Study of Lung Cancer (IASLC) to Dr. Pilotto. The funding agencies had no role in the collection, analysis, and interpretation of data or in the writing of the article. Dr. Simbolo, Dr. Bria, and Dr. Scarpa conceived the study. Dr. Simbolo designed the study and validation experiments. Dr. Vicentini supervised the validation experiments. Dr. Lawlor coordinated patients and sample data management and supervised ethical protocols. Dr. Ali, Dr. Pilotto, Dr. Mastracci, Dr. Grillo, Dr. Fontanini, Dr. Volante, and Dr. Bria collected materials and clinical data. Dr. Ali, Dr. Fassan, Dr. Fontanini, Dr. Volante, Dr. Pelosi, Dr. Scarpa analyzed histopathologic data. Dr. Rusev and Dr. Fassan microdissected samples. Ms. Sperandio and Dr. Pelliccioni extracted and qualified nucleic acids. Dr. Simbolo, Dr. Mafficini, and Dr. Barbi carried out sequencing and performed bioinformatic analysis. Dr. Ali, Dr. Mastracci, and Dr. Vicentini performed the immunohistochemistry analysis. Dr. Simbolo, Dr. Barbi, Dr. Mafficini, Dr. Fassan, Dr. Corbo, Dr. Tortora, Dr. Volante, Dr. Pelosi drafted the article. Dr. Scarpa and Dr. Bria revised and finalized the article. All authors approved the submitted version.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at www.jto.org and at https://doi.org/10.1016/j.jtho.2019.05.003.

References

1. Travis WD, Brambilla E, Burke AP, Marx Z, Nicholson AG, eds. *WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart*. 4th ed. Lyon, France: IARC; 2015.
2. Asamura H, Kameya T, Matsuno Y, et al. Neuroendocrine neoplasms of the lung: a prognostic spectrum. *J Clin Oncol*. 2006;24:70–76.
3. Beasley MB, Thunnissen FB, Brambilla E, et al. Pulmonary atypical carcinoid: predictors of survival in 106 cases. *Hum Pathol*. 2000;31:1255–1265.
4. Swarts DR, Ramaekers FC, Speel EJ. Molecular and cellular biology of neuroendocrine lung tumors: evidence for separate biological entities. *Biochim Biophys Acta*. 2012;1826:255–271.
5. Fernandez-Cuesta L, Peifer M, Lu X, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat Commun*. 2014;5:3518.
6. Clinical Lung Cancer Genome Project (CLCGP), Network Genomic Medicine (NGM). A genomics-based classification of human lung tumors. *Sci Transl Med*. 2013;5:209ra153.
7. Derks JL, Leblay N, Lantuejoul S, Dingemans AC, Speel EM, Fernandex-Cuesta L. New insights into the

molecular characteristics of pulmonary carcinoids and large cell neuroendocrine carcinomas, and the impact on their clinical management. *J Thorac Oncol*. 2018;13:752–766.
8. Simbolo M, Mafficini A, Sikora KO, et al. Lung neuroendocrine tumours: deep sequencing of the four World Health Organization histotypes reveals chromatin-remodelling genes as major players and a prognostic role for TERT, RB1, MEN1 and KMT2D. *J Pathol*. 2017;241:488–500.
9. George J, Walter V, Peifer M, et al. Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nat Commun*. 2018;9:1048.
10. Umemura S, Mimaki S, Makinoshima H, et al. Therapeutic priority of the PI3K/AKT/mTOR pathway in small cell lung cancers as revealed by a comprehensive genomic analysis. *J Thorac Oncol*. 2014;9:1324–1331.
11. Vollbrecht C, Werner R, Walter RF, et al. Mutational analysis of pulmonary tumours with neuroendocrine features using targeted massive parallel sequencing: a comparison of a neglected tumour group. *Br J Cancer*. 2015;113:1704–1711.
12. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*. 2012;150:1121–1134.
13. Pelosi G, Bianchi F, Dama E, et al. Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. *Virchows Arch*. 2018;472:567–577.
14. Karlsson A, Brunnstrom H, Micke P, et al. Gene expression profiling of large cell lung cancer links transcriptional phenotypes to the new histological WHO 2015 classification. *J Thorac Oncol*. 2017;12:1257–1267.
15. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*. 2001;98:13790–13795.
16. Asiedu MK, Thomas CF Jr, Dong J, et al. Pathways impacted by genomic alterations in pulmonary carcinoid tumors. *Clin Cancer Res*. 2018;24:1691–1704.
17. Travis WD. Advances in neuroendocrine lung tumors. *Ann Oncol*. 2010;21(suppl 7):vii65–vii71.
18. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. 2015;10:1243–1260.
19. Travis WD, Rush W, Flieder DB, et al. Survival analysis of 200 pulmonary neuroendocrine tumors with clarification of criteria for atypical carcinoid and its separation from typical carcinoid. *Am J Surg Pathol*. 1998;22:934–944.
20. Rekhtman N. Neuroendocrine tumors of the lung: an update. *Arch Pathol Lab Med*. 2010;134:1628–1638.
21. Vallieres E, Shepherd FA, Crowley J, et al. The IASLC Lung Cancer Staging Project: proposals regarding the relevance of TNM in the pathologic staging of small cell lung cancer in the forthcoming (seventh) edition of the TNM classification for lung cancer. *J Thorac Oncol*. 2009;4:1049–1059.
22. Simbolo M, Gottardi M, Corbo V, et al. DNA qualification workflow for next generation sequencing of histopathological samples. *PloS One*. 2013;8:e62692.
23. Zamo A, Bertolaso A, van Raaij AW, et al. Application of microfluidic technology to the BIOMED-2 protocol for detection of B-cell clonality. *J Mol Diagn*. 2012;14:30–37.
24. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet*. 2012;3:35.
25. McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26:2069–2070.
26. Robinson JT, Thorvaldsdottir H, Winckler W, Zehir A, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011;29:24–26.
27. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415–421.
28. Diaz-Gay M, Vila-Casadesus M, Franch-Exposito S, Hernández-Ilián E, Lozano JJ, Castellvi-Bel S. Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics*. 2018;19:224.
29. Boeva V, Popova T, Lienard M, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*. 2014;30:3443–3450.
30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
31. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.
32. R: a language and environment for statistical computing [computer program]. Vienna, Austria: 2015.
33. Chen HZ, Tsai SY, Leone G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer*. 2009;9:785–797.
34. Swarts DR, Scarpa A, Corbo V, et al. MEN1 gene mutation and reduced expression are associated with poor prognosis in pulmonary carcinoids. *J Clin Endocrinol Metab*. 2014;99:E374–E378.
35. Rekhtman N, Pietanza MC, Hellmann M, et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma-like and non-small cell carcinoma-like subsets. *Clin Cancer Res*. 2016;22:3618–3629.
36. Corbo V, Dalai I, Scardoni M, et al. MEN1 in pancreatic endocrine tumors: analysis of gene and protein status in 169 sporadic neoplasms reveals alterations in the vast majority of cases. *Endocr Relat Cancer*. 2010;17:771–783.
37. Scarpa A, Chang DK, Nones K, et al. Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature*. 2017;543:65–71.
38. Alcala N, Leblay N, Gabriel A, et al. Integrative and comparative genomic analyses identify clinically relevant groups of pulmonary carcinoids and unveil the existence of supra-carcinoids. Paper presented at: 16th Annual European Neuroendocrin Tumor Society Conference. March 6–8, 2019; Barcelona, Spain.

HHS Public Access

Author manuscript

Nat Commun. Author manuscript; available in PMC 2014 September 27.

Published in final edited form as:

Nat Commun. ; 5: 3518. doi:10.1038/ncomms4518.

Frequent mutations in chromatin-remodeling genes in pulmonary carcinoids

Lynnette Fernandez-Cuesta[#1], **Martin Peifer**[#1,2], **Xin Lu**[1], **Ruping Sun**[3], **Luka Ozretić**[4], **Danila Seidal**[1,5], **Thomas Zander**[1,6,7], **Frauke Leenders**[1,5], **Julie George**[1], **Christian Müller**[1], **Ilona Dahmen**[1], **Berit Pinther**[1], **Graziella Bosco**[1], **Kathryn Konrad**[8], **Janine Altmüller**[8,9,10], **Peter Nürnberg**[2,8,9], **Viktor Achter**[11], **Ulrich Lang**[11,12], **Peter M Schneider**[13], **Magdalena Bogus**[13], **Alex Soltermann**[14], **Odd Terje Brustugun**[15,16], **Åslaug Helland**[15,16], **Steinar Solberg**[17], **Marius Lund-Iversen**[18], **Sascha Ansén**[6], **Erich Stoelben**[19], **Gavin M. Wright**[20], **Prudence Russell**[21], **Zoe Wainer**[20], **Benjamin Solomon**[22], **John K Field**[23], **Russell Hyde**[23], **Michael PA. Davies**[23], **Lukas C Heukamp**[4,7], **Iver Petersen**[24], **Sven Perner**[25], **Christine Lovly**[26], **Federico Cappuzzo**[27], **William D Travis**[28], **Jürgen Wolf**[5,6,7], **Martin Vingron**[3], **Elisabeth Brambilla**[29], **Stefan A. Haas**[3], **Reinhard Buettner**[4,5,7], and **Roman K Thomas**[1,4,5]

[1]Department of Translational Genomics, Center of Integrated Oncology Cologne–Bonn, University of Cologne, 50924 Cologne, Germany [2]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany [3]Computational Molecular Biology Group, Max Planck Institute for Molecular Genetics, D-14195 Berlin, Germany [4]Department of Pathology, University Hospital Medical Center, University of Cologne, 50937 Cologne, Germany [5]Laboratory of Translational Cancer Genomics, Center of Integrated Oncology Cologne – Bonn, University of Cologne, 50924 Cologne, Germany [6]Department I of Internal Medicine, Center of Integrated Oncology Kö ln-Bonn, University of Cologne, 50924 Cologne, Germany [7]Network Genomic Medicine, University Hospital Cologne, Center of Integrated Oncology Cologne Bonn, 50924 Cologne, Germany [8]Cologne Center for Genomics (CCG), University of Cologne, Cologne, 50931

Correspondence: Roman K Thomas, Department of Translational Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany, +49-221-478-98771, roman.thomas@uni-Köln.de.

Author contributions

LFC and RKT conceived the project. LFC, MP and RKT analyzed, interpreted the data, and wrote the manuscript. LO, CM, ID, BP, KK, JA, and MB performed experiments. LFC, MP and XL performed computational analysis. MP, RS and SAH provided unpublished algorithms. LFC, MP, TZ, RB and RKT gave scientific input. AS, OTB, AH, SS, MLI, SA, ES, GMW, PR, ZW, BS, JKF, RH, MPAD, LCH, IP, SP, CL, FC, EB and RB contributed with samples. LO, WDT, EB, and RB performed pathology review. DS, FL, JG, GB, PN, VA, UL, PMS, SA, JW and MV helped with logistics. All the co-authors reviewed the manuscript.

Competing financial interests

RKT is a founder and shareholder of Blackfield AG. RKT received consulting and lecture fees (Sanofi- Aventis, Merck, Roche, Lilly, Boehringer Ingelheim, Astra-Zeneca, Atlas-Biolabs, Daiichi-Sankyo, MSD, Blackfield AG, Puma) as well as research support (Merck, EOS and AstraZeneca). RB is a cofounder and – owner of Targos Molecular Diagnostics and received honoraria for consulting and lecturing from AstraZeneca, Boehringer Ingelheim, Merck, Roche, Novartis, Lilly, and Pfizer. JW received consulting and lecture fees from Roche, Novartis, Boehringer Ingelheim, AstraZeneca, Bayer, Lilly, Merck, Amgen and research support from Roche, Bayer, Novartis, Boehringer Ingelheim. TZ received honoraria from Roche, Novartis, Boehringer Ingelheim, Lilly, Merck, Amgen and research support from Novartis. CML has served on an Advisory Board for Pfizer and has served as a speaker for Abbott and Qiagen. The remaining authors declare no competing financial interests.

Accession Codes

Sequence data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001000650.

Germany [9]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany [10]Institute of Human Genetics, University of Cologne, Cologne 50931, Germany [11]Computing Center, University of Cologne, 50931 Cologne, Germany [12]Department of Informatics, University of Cologne, 50931 Cologne, Germany [13]Institute of Legal Medicine, University of Cologne, 50823 Cologne, Germany [14]Institute for Surgical Pathology, University Hospital Zurich, 8091 Zurich, Switzerland [15]Institute of clinical medicine, Faculty of Medicine, University of Oslo, N-0424 Oslo, Norway [16]Department of oncology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway [17]Department of Thoracic Surgery, Rikshospitalet, Oslo University Hospital, N-0027 Oslo, Norway [18]Department of pathology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway [19]Thoracic Surgery, Lungenklinik Merheim, Kliniken der Stadt Köln gGmbH, 51109 Cologne, Germany [20]University of Melbourne Department of Surgery, St Vincent's Hospital, Melbourne, 3065 Victoria, Australia [21]Department of Pathology, St. Vincent's Hospital, Melbourne, 3065 Victoria, Australia [22]Department of Haematology and Medical Oncology, Peter MacCallum Cancer Centre, Melbourne, 3002 Victoria, Australia [23]Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, University of Liverpool Cancer Research Centre, Liverpool, L3 9TA, UK [24]Institute of Pathology, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany [25]Department of Prostate Cancer Research, Institute of Pathology, University Hospital of Bonn, 53127 Bonn, Germany [26]Vanderbilt-Ingram Cancer Center, Nashville, TN37232, USA [27]Department of Medical Oncology, Istituto Toscano Tumouri, 57100 Livorno, Italy [28]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York 10065, USA [29]Department of Pathology, CHU Grenoble INSERM U823, Institute Albert Bonniot 38043 CS10217 Grenoble, France

[#] These authors contributed equally to this work.

Abstract

Pulmonary carcinoids are rare neuroendocrine tumors of the lung. The molecular alterations underlying the pathogenesis of these tumors have not been systematically studied so far. Here we perform gene copy number analysis (n=54), genome/exome (n=44) and transcriptome (n=69) sequencing of pulmonary carcinoids and observe frequent mutations in chromatin-remodeling genes. Covalent histone modifiers and subunits of the SWI/SNF complex are mutated in 40% and 22.2% of the cases respectively, with *MEN1*, *PSIP1* and *ARID1A* being recurrently affected. In contrast to small-cell lung cancer and large-cell neuroendocrine tumors, *TP53* and *RB1* mutations are rare events, suggesting that pulmonary carcinoids are not early progenitor lesions of the highly aggressive lung neuroendocrine tumors but arise through independent cellular mechanisms. These data also suggest that inactivation of chromatin remodeling genes is sufficient to drive transformation in pulmonary carcinoids.

Introduction

Pulmonary carcinoids are neuroendocrine tumors that account for about 2% of pulmonary neoplasms. Based on the WHO classification of 2004, carcinoids can be subdivided in

typical or atypical, the latter ones being very rare (about 0.2%)[1]. Most carcinoids can be cured by surgery; however, inoperable tumors are mostly insensitive to chemo- and radiation therapies1. Apart from few low-frequency alterations, such as mutations in *MEN1*[1], comprehensive genome analyses of this tumor type have so far been lacking.

Here we conduct integrated genome analyses[2] on data from chromosomal gene copy number of 54 tumors, genome and exome sequencing of 29 and 15 tumor-normal pairs respectively, as well as transcriptome sequencing of 69 tumors. Chromatin-remodeling is the most frequently mutated pathway in pulmonary carcinoids; the genes *MEN1*, *PSIP1* and *ARID1A* were recurrently affected by mutations. Specifically, covalent histone modifiers and subunits of the SWI/SNF complex are mutated in 40% and 22.2% of the cases respectively. By contrast, mutations of *TP53* and *RB1* are only found in 2 out of 45 cases, suggesting that these genes are not main *drivers* in pulmonary carcinoids.

Results

In total, we generated genome/exome sequencing data for 44 independent tumor-normal pairs, and for most of them, also RNAseq (n=39, 69 in total), and SNP 6.0 (n=29, 54 in total) data (**Supplementary Table S1)**. Although no significant focal copy number alterations were observed across the tumors analyzed, we detected a copy number pattern compatible with chromothripsis3 in a stage-III atypical carcinoid of a former smoker (**Fig. 1a; Supplementary Fig. S1**). The intensely clustered genomic structural alterations found in this sample were restricted to chromosomes 3, 12, and 13, and led to the expression of several chimeric transcripts (**Fig. 1b; Supplementary Table S2**). Some of these chimeric transcripts affected genes involved in chromatin remodeling processes, including out-of-frame fusion transcripts disrupting the genes, *ARID2*, *SETD1B*, and *STAG1*. Through the analyses of genome and exome sequencing data, we detected 529 non-synonymous mutations in 494 genes, which translates to a mean somatic mutation rate of 0.4 mutations per megabase (Mb) (**Fig. 1c; Supplementary Data 1**), which is much lower than the rate observed in other lung tumors (**Fig. 1c**)[2,4,5]. As expected, and in contrast to small-cell lung cancer (SCLC), no smoking-related mutation signature was observed in the mutation pattern of pulmonary carcinoids (**Fig. 1d**).

We identified *MEN1*, *ARID1A* and *EIF1AX* as significantly mutated genes[2] (q-value<0.2, see Methods section) (Fig. 2a; Supplementary Table S1 and S3; Supplementary Data 1). *MEN1* and *ARID1A* play important roles in chromatin remodeling processes. The tumor suppressor MEN1 physically interacts with MLL and MLL2 to induce gene transcription[6]. Specifically, MEN1 is a molecular adaptor that physically links MLL with the chromatin-associated protein PSIP1, an interaction that is required for MLL/MEN1-dependent functions7. MEN1 also acts as a transcriptional repressor through the interaction with SUV39H1[8]. We observed mutually exclusive frame-shift and truncating mutations in *MEN1* and *PSIP1* in 6 cases (13.3%), which were almost all accompanied by loss of heterozygosity (LOH) (**Supplementary Fig. S2**). We also detected mutations in histone methyltransferases (*SETD1B*, *SETDB1* and *NSD1*) and demethylases (*KDM4A*, *PHF8* and *JMJD1C*), as well as in the following members of the Polycomb complex[9] (Supplementary Table S1 and S2; Supplementary Data 1): *CBX6*, which belongs to the Polycomb repressive complex 1

(PRC1); *EZH1*, which is part of the Polycomb repressive complex 2 (PRC2); and *YY1*, a member of the PHO repressive complex 1 that recruits PRC1 and PRC2. *CBX6* and *EZH1* mutations were also accompanied by LOH (**Supplementary Fig. S2**). In addition, we also detected mutations in the histone modifiers *BRWD3* and *HDAC5* in one sample each. In total, 40% of the cases carried mutually exclusive mutations in genes that are involved in covalent histone modifications (*q*-value=$8x10^{-7}$, see Methods section) (**Fig. 2a; Supplementary Table S4**). In order to evaluate the impact of these mutations on histone methylation, we compared the levels of the H3K9me3 and H3K27me3 on 7 mutated and 6 wild-type samples, and observed a trend towards lower methylation in the mutated cases (**Table 1; Fig. 2b**).

Truncating and frame-shift mutations in *ARID1A* were detected in 3 cases (6.7%). ARID1A is one of the two mutually exclusive ARID1 subunits, believed to provide specificity to the ATP-dependent SWI/SNF chromatin-remodeling complex[10,11]. Truncating mutations of this gene have been reported at high frequency in several primary human cancers[12]. In total, members of this complex were mutated in mutually exclusive fashion in 22.2% of the specimens (*q*-value=$8x10^{-8}$, see Methods section) (**Fig. 2a; Supplementary Table S4**). Among them were the core subunits *SMARCA1*, *SMARCA2*, and *SMARCA4*, which carry the ATPase activity of the complex, as well as the subunits *ARID2*, *SMARCC2*, *SMARCB1*, and, *BCL11A* (Fig. 2a; Supplementary Table S1 and S2; Supplementary Data 1)[13,14]. Another recurrently affected pathway was sister-chromatid cohesion during cell cycle progression with the following genes mutated (Fig. 2a; Supplementary Table S1 and S2; Supplementary Data 1; Supplementary Fig. S3): the cohesin subunit *STAG1*[15], the cohesin loader *NIPBL*[16]; the ribonuclease and microRNA processor *DICER*, necessary for centromere establishment[17]; and ERCC6L, involved in sister chromatid separation[18]. In addition, although only few chimeric transcripts were detected in the 69 transcriptomes analyzed (**Supplementary Table S5),** we found one sample harboring an inactivating chimeric transcript leading to the loss of the mediator complex gene *MED24* (**Supplementary Fig. S4**) that interacts both physically and functionally with cohesin and NIPBL to regulate gene expression[19]. In summary, we detected mutations in chromatin remodeling genes in 23 (51.1%) of the samples analyzed. The specific role of histone modifiers in the development of pulmonary carcinoids was confirmed by the lack of significance of these pathways in SCLC[2] (**Supplementary Table S4**). This was further supported by a gene expression analysis including 50 lung adenocarcinomas (unpublished data), 42 SCLC[2,20], and the 69 pulmonary carcinoids included in this study (**Supplementary Data 2**). Consensus k-means clustering revealed that although both SCLC and pulmonary carcinoids are lung neuroendocrine tumors, both tumor types as well as adenocarcinomas formed statistically significant separate clusters (**Fig. 3a**). In support of this notion, we recently reported that the early alterations in SCLC universally affect *TP53* and *RB1*[2], whereas in this study these genes were only mutated in two samples (Fig 2a; Supplementary Table S1; Supplementary Data 1). Moreover, when examining up- and down-regulated pathways in SCLC versus pulmonary carcinoids by Gene Set Enrichment Analysis (GSEA)[21], we found that in line with the pattern of mutations, the RB1 pathway was statistically significantly altered in SCLC (q-value=$5x10^{-4}$, see Methods section) but not in pulmonary carcinoids (**Fig. 3b; Supplementary Table S6**).

Another statistically significant mutated gene was the eukaryotic translation initiation factor 1A (*EIF1AX*) mutated in 4 cases (8.9%). Additionally, *SEC31A*, *WDR26*, and the E3-ubiquitin ligase *HERC2* were mutated in two samples each. Further supporting a role of E3 ubiquitin ligases in the development of pulmonary carcinoids we found mutations or rearrangements affecting these genes in 17.8% of the samples analyzed (Fig. 2a; Supplementary Table S1 and S7; Supplementary Data 1). All together, we have identified candidate driver genes in 73.3% of the cases. Of note, we did not observe any genetic segregation between typical or atypical carcinoids, neither between the expression clusters generated from the two subtypes, nor between these clusters and the mutated pathways (**Supplementary Fig. S5**). However, it is worth mentioning that only 9 atypical cases were included in this study. The spectrum of mutations found in the discovery cohort, was further validated by transcriptome sequencing of an independent set of pulmonary carcinoid specimens (**Supplementary Table S1 and S8**). Due to the fact that many nonsense and frame-shift mutations may result in nonsense-mediated decay[22,23], the mutations detected by transcriptome sequencing were only missense. Due to this bias, accurate mutation frequencies could not be inferred from these data.

Discussion

This study defines recurrently mutated sets of genes in pulmonary carcinoids. The fact that almost all of the reported genes were mutated in a mutually exclusive manner and affected a small set of cellular pathways, defines these as the key pathways in this tumor type. Given the frequent mutations affecting the few signaling pathways described above and the almost universal absence of other cancer mutations, our findings support a model where pulmonary carcinoids are not early progenitor lesions of other neuroendocrine tumors, such as small-cell lung cancer or large-cell neuroendocrine carcinoma, but arise through independent cellular mechanisms. More broadly, our data suggest that mutations in chromatin remodeling genes, which in recent studies were found frequently mutated across multiple malignant tumours[24], are sufficient to drive early steps in tumorigenesis in a precisely defined spectrum of required cellular pathways.

Methods

Tumor specimens

The study as well as written informed consent documents had been approved by the Institutional Review Board of the University of Cologne. Additional biospecimens for this study were obtained from the Victorian Cancer Biobank, Melbourne, Australia; the Vanderbilt-Ingram Cancer Center, Nashville, USA; and Roy Castle Lung Cancer Research Programme, The University of Liverpool Cancer Research Center, Liverpool, UK. The Institutional Review Board (IRB) of each participating institution approved collection and use of all patient specimens in this study.

Nucleic acid extraction and sample sequencing

All samples in this study were reviewed by expert pathologists. Total RNA and DNA were obtained from fresh-frozen tumor and matched fresh-frozen normal tissue or blood. Tissue

was frozen within 30 min after surgery and was stored at –80 °C. Blood was collected in tubes containing the anticoagulant EDTA and was stored at –80 °C. Total DNA and RNA were extracted from fresh-frozen lung tumor tissue containing more than 70% tumor cells. Depending on the size of the tissue, 15–30 sections, each 20 μm thick, were cut using a cryostat (Leica) at –20 °C. The matched normal sample obtained from frozen tissue was treated accordingly. DNA from sections and blood was extracted using the Puregene Extraction kit (Qiagen) according to the manufacturer's instructions. DNA was eluted in 1× TE buffer (Qiagen), diluted to a working concentration of 150 ng—l and stored at –80 °C. For whole exome sequencing we fragmented 1 μg of DNA with sonification technology (Bioruptor, diagenode, Liège, Belgium). The fragments were endrepaired and adaptor-ligated, including incorporation of sample index barcodes. After size selection, we subjected the library to an enrichment process with the SeqCap EZ Human Exome Library version 2.0 kit (Roche NimbleGen, Madison, WI, USA). The final libraries were sequenced with a paired-end 2×100 bp protocol. On average, 7 Gb of sequence were produced per normal, resulting in 30x coverage of more than 80% of target sequences (44Mb). For better sensitivity, tumors were sequenced with 12Gb and 30x coverage of more than 90%. We filtered primary data according to signal purity with the Illumina Realtime Analysis software. Whole genome sequencing was also performed using a read length of 2x 100bp for all samples. On average, 110 Gb of sequence were produced per sample, aiming a mean coverage of 30x for both tumor and matched-normal. RNAseq was performed on cDNA libraries prepared from PolyA+ RNA extracted from tumor cells using the Illumina TruSeq protocol for mRNA. The final libraries were sequenced with a paired-end 2×100 bp protocol aiming at 8.5 Gb per sample, resulting on a 30x mean coverage of the annotated transcriptome. All the sequencing was carry on an Illumina HiSeq™ 2000 sequencing instrument (Illumina, San Diego, CA, USA).

Sequence data processing and mutation detection

Raw sequencing data are aligned to the most recent build of the human genome (NCBI build 37/hg19) using BWA (version: 0.5.9rc1)[25] and possible PCR-duplicates are subsequently removed form the alignments. Somatic mutations were detected using our in-house developed sequencing analysis pipeline. In brief, the mutation calling algorithm incorporates parameters such as local copy number profiles, estimates of tumor purity and ploidy, local sequencing depth, as well as the global sequencing error into a statistical model with which the presence of a mutated allele in the tumor is determined. Next, the absence of this variant in the matched normal is assessed by demanding that the corresponding allelic fraction is compatible with the estimated background sequencing error in the normal. In addition, we demand that the allelic fractions between tumor and normal differ significantly. To finally remove artificial mutation calls, we apply a filter that is based on the forward-reverse bias of the sequencing reads. Further details of this approach are given in Peifer *et* al.[2]

Genomic rearrangement reconstruction from paired-end data

To reconstruct rearrangements from paired-end data, we refined our initial method[2] by adding breakpoint-spanning reads. Here, locations of encompassing read pairs are screened for further reads where only one pair aligns to the region and the other pair either does not align at all or is clipped by the aligner. These reads are then realigned using BLAT to a

1000bp region around the region defined by the encompassing reads. Rearrangements confirmed by at least one spanning read are finally reported. To filter for somatic rearrangements, we subtracted those regions where rearrangements are present in the matched-normal and in all other sequenced normals within the project.

Analysis of significantly mutated genes and pathways

The analysis of significantly mutated genes is done in a way that both gene expression and the accumulation of synonymous mutations are considered to obtain robust assessments of frequently mutated, yet biologically relevant genes. To this end, the overall background mutation rate is determined first, from which the expected number of mutations for each gene is computed under the assumption of a purely random mutational process. This gene specific expected number of mutations defines the underlying null model of our statistical test. To account for misspecifications, e.g., due to a local variation of mutation rates, we also incorporated the synonymous to non-synonymous ratio into a combined statistical model to determine significantly mutated genes. Since mutation rates in non-expressed genes are often high than the genome-wide background rate[2,26], genes that are having a median FPKM value less than one in our transcriptome sequencing data are removed prior testing. To account for multiple hypothesis testing, we are using the Benjamini-Hochberg approach[27]. Mutation data of the total of 44 samples, for which either WES or WGS was performed, were used for this analysis.

In case of the pathway analysis, gene lists of the methylation- and the SWI/SNF complex were obtained from recent publications[9,13,14,28]. To assess whether mutations in these pathways are significantly enriched, all genes of the pathway are grouped together as if they represent a ”single gene” and subsequently tested if the total number of mutation exceed mutational background of the entire pathway. To this end, the same method as described above was used. Mutation data of the total of 44 samples, for which either WES or WGS was performed, were used for this analysis.

Analysis of chromosomal gene copy number data

Hybridization of the Affymetrix SNP 6.0 arrays was carried out according to the manufacturers' instructions and analyzed as follows: raw signal intensities were processed by applying a log-linear model to determine allele-specific probe affinities and probe-specific background intensities. To calibrate the model, a Gauss-Newton approach was used and the resulting raw copy number profiles are segmented by applying the circular binary segmentation method[29].

Analysis of RNAseq data

For the analysis of RNAseq data, we have developed a pipeline that affords accurate and efficient mapping and downstream analysis of transcribed genes in cancer samples (Lynnette Fernandez-Cuesta and Ruping Sun, personal communication). In brief, paired-end RNAseq reads were mapped onto hg19 using a sensitive gapped aligner, GSNAP[30]. Possible breakpoints were called by identifying individual reads showing split-mapping to distinct locations as well as clusters of discordant read pairs. Breakpoint assembly was performed to

leverage information across reads anchored around potential breakpoints. Assembled contigs were aligned back to the reference genome to confirm *bona fide* fusion points.

Dideoxy sequencing

All non-synonymous mutations found in the genome/exome data were checked in RNAseq data when available. Genes recurrently mutated involved in pathways statistically significantly mutated, or interesting because of their presence in other lung neuroendocrine tumors, were selected for validation. 158 mutations were considered for validation: 115 validated and 43 did not (validation rate 73%). Sequencing primer pairs were designed to enclose the putative mutation (**Supplementary Data 1**), or to encompass the candidate rearrangement (**Supplementary Table S7**) or chimeric transcript (**Supplementary Table S2 and S5**). Sequencing was carried out using dideoxy-nucleotide chain termination (Sanger) sequencing, and electropherograms were analyzed by visual inspection using 4 Peaks.

Gene expression data analyses

Unsupervised consensus clustering was applied to RNAseq data of 69 pulmonary carcinoids, 50 AD, and 42 SCLC[2,20] samples. The 3000 genes with highest variation across all samples were filtered out before performing consensus clustering. We used the clustering module from GenePattern[31] and the consensus CDF[32,33]. Significance was obtained by using SigClust[34]. Fisher's exact test[35] was used to check for associations between clusters and histological subtypes. GSEA[21] were performed on 69 pulmonary carcinoids and 42 SCLC[2,20] samples; and the gene sets *oncogenic signatures* were used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are indebted to the patients donating their tumor specimens as part of the Clinical Lung Cancer Genome Project initiative. We thank Philipp Lorimier, Elisabeth Kirst, Emilia Müller, and Juana Cuesta Valdes for their technical assistance. We furthermore thank the regional computing center of the University of Cologne (RRZK) for providing the CPU time on the DFG-funded supercomputer 'CHEOPS' as well as the support.

This work was supported by the Deutsche Krebshilfe as part of the small-cell lung cancer genome-sequencing consortium (grant ID: 109679 to RKT, MP, RB, PN, MV and SAH). Additional funds were provided by the EU-Framework program CURELUNG (HEALTH-F2-2010-258677 to RKT, JW, JKF and EB); by the German federal state North Rhine Westphalia (NRW) and the European Union (European Regional Development Fund: *Investing In Your Future*) within PerMed NRW (grant 005-1111-0025 to RKT, JW, RB); by the Deutsche Forschungsgemeinschaft through TH1386/3-1 (to RKT) and through SFB832 (TP6 to RKT and JW; TP5 to LCH); by the German Ministry of Science and Education (BMBF) as part of the NGFNplus program (grant 01GS08101 to RKT, JW, PN); by the Deutsche Krebshilfe as part of the *Oncology Centers of Excellence* funding program (RKT, RB, JW); by *Stand Up To Cancer*–American Association for Cancer Research Innovative Research Grant (SU2C-AACR-IR60109 to RKT); by an NIH K12 training grant (K12 CA9060625) and by an *Uniting Against Lung Cancer* grant, and a Damon Runyon Clinical Investigator Award (to CML); and by AIRC and Istituto Toscano Tumori project F13/16 (to FC).

References

1. Swarts, D. R. a; Ramaekers, FCS.; Speel, E-JM. Molecular and cellular biology of neuroendocrine lung tumors: evidence for separate biological entities. Biochim. Biophys. Acta. 2012; 1826:255–271. [PubMed: 22579738]
2. Peifer M, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat. Genet. 2012; 44:1104–1110. [PubMed: 22941188]
3. Stephens PJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011; 144:27–40. [PubMed: 21215367]
4. Imielinski M, et al. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. Cell. 2012; 150:1107–1120. [PubMed: 22980975]
5. Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012
6. Marx SJ. Molecular genetics of multiple endocrine neoplasia types 1 and 2. Nat. Rev. Cancer. 2005; 5:367–376. [PubMed: 15864278]
7. Yokoyama A, Clearly M. Menin critically links MLL proteins with LEDGF on cancer-associated target genes. Cancer Cell. 2008; 8:2469.
8. Yang Y-J, et al. Menin mediates epigenetic regulation via histone H3 lysine 9 methylation. Cell Death & Disease. 2013; 4:e583. [PubMed: 23579270]
9. Lanzuolo C, Orlando V. Memories from the polycomb group proteins. Annu. Rev. Genet. 2012; 46:561–89. [PubMed: 22994356]
10. Roberts CWM, Orkin SH. The SWI/SNF complex — chromatin and cancer. Nat. Rev. Cancer. 2004; 4:133–142. [PubMed: 14964309]
11. Wu JI, Lessard J, Crabtree GR. Understanding the words of chromatin regulation. Cell. 2009; 136:200–206. [PubMed: 19167321]
12. Wilson BG, Roberts CWM. Epigenetics and genetics SWI/SNF nucleosome remodellers and cancer. Nat. Rev. Cancer. 2011; 11:481–492. [PubMed: 21654818]
13. Tang J, Yoo AS, Crabtree GR. Reprogramming human fibroblasts to neurons by recapitulating an essential microRNA-chromatin switch. Current opinion in genetics & development. 2013; 23:591–8. [PubMed: 24035011]
14. Kadoch C, et al. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. 2013; 45:1–11. [PubMed: 23268125]
15. Peters J, Tedeschi A, Schmitz J. The cohesin complex and its roles in chromosome biology. Genes Dev. 2008; 22:3089–3114. [PubMed: 19056890]
16. Ciosk R, et al. Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins. Mol Cell. 2000; 5:243–254. [PubMed: 10882066]
17. Fukagawa T, et al. Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat. Cell Biol. 2004; 6:784–791. [PubMed: 15247924]
18. Baumann C, Korner R, Hofmann K, Nigg EA. PICH, a centromere-associated SNF2 family ATPase, is regulated by Plk1 and required for the spindle checkpoint. Cell. 2007; 128:101–114. [PubMed: 17218258]
19. Kagey MH, et al. Mediator and Cohesin Connect Gene Expression and Chromatin Architecture. Nature. 2010; 467:430–435. [PubMed: 20720539]
20. Rudin CM, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat. Genet. 2012; 44:1111–1116. [PubMed: 22941189]
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005
22. Nicholson P, et al. Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. Cellular and molecular life sciences: CMLS. 2010; 67:677–700. [PubMed: 19859661]
23. Yepiskoposyan H, Aeschimann F, Nilsson D, Okoniewski M, Muhlemann O. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. RNA (New York, N.Y.). 2011; 17:2108–18.

24. Timp W, Feinberg AP. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nature Reviews Cancer. 2013; 13:497–510.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009; 25:1754–60.
26. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013:10–14.
27. 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). 1995; 57
28. Black JC, Van Rechem C, Whetstine JR. Histone lysine methylation dynamics: establishment, regulation, and biological impact. Mol. Cell. 2012; 48:491–507. [PubMed: 23200123]
29. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics (Oxford, England). 2004; 5:557–72.
30. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics (Oxford, England). 2010; 26:873–81.
31. Kuehn H, Liberzon A, Reich M, Mesirov JP. Using GenePattern for gene expression analysis. Curr. Protoc. Bioinformatics. 2008; 22:7.12.1–7.12.39.
32. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics (Oxford, England). 2010; 26:1572–3.
33. Monti S, et al. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning Journal. 2003; 52(1-2):91–118.
34. Liu Y, et al. Statistical significance of clustering for high-dimension, low-sample size data. J Am. Stat. Assoc. 2008; 103:1281–1293.
35. Fisher RA. Statistical Methods for Research Workers. Oliver and Boyd. 1954
36. Karro JE, et al. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. Mol. Biol. Evol. 2008; 25:362–374. [PubMed: 18042807]

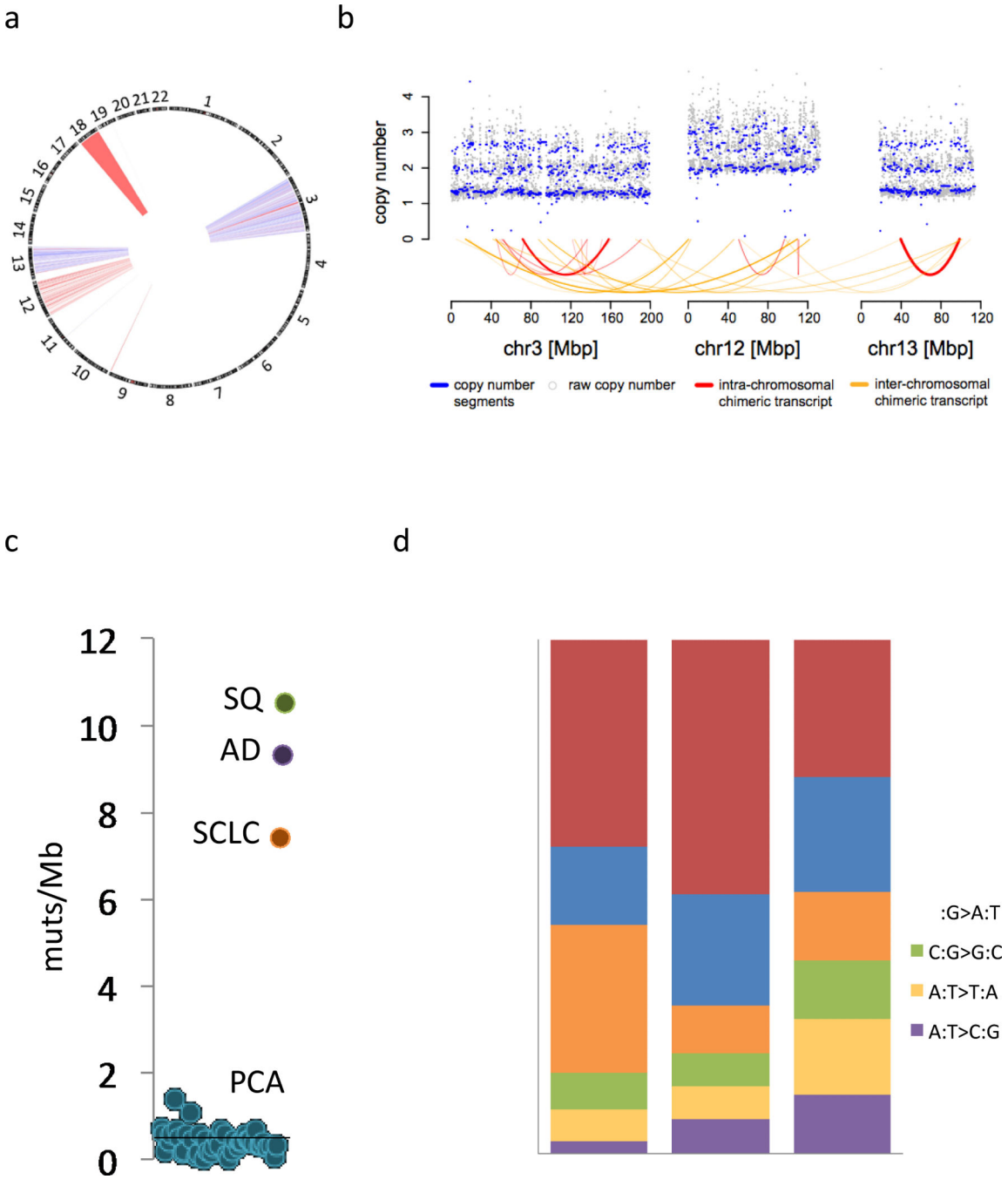


Figure 1.

Genomic characterization of pulmonary carcinoids. (a) CIRCOS plot of the chromothripsis case. The outer ring shows chromosomes arranged end to end. Somatic copy number alterations (gains in red and losses in blue) detected by 6.0 SNP arrays are depicted in the inside ring. (**b**) Copy numbers and chimeric transcripts of affected chromosomes. Segmented copy number states (blue points) are shown together with raw copy number data averaged over 50 adjacent probes (grey points). To show the different levels of strength for the identified chimeric transcripts all curves are scaled according to the sequencing coverage

at the fusion-point. (**c**) Mutation frequency detected by genome and exome sequencing in pulmonary carcinoids (PCA). Each blue dot represents the number of mutations per megabase in one pulmonary carcinoid sample. Average frequencies are also shown for adenocarcinomas (AD), squamous (SQ), and small-cell lung cancer (SCLC) base on previous studies[2,4,5] (**d**) Comparison of context independent transversion and transition rates (an overall strand symmetry is assumed) between rates derived from molecular evolution (evol)[36], from a previous SCLC sequencing study[2], and from the pulmonary carcinoids (PCA) genome and exome sequencing. All rates are scaled as such that their overall sum is one.

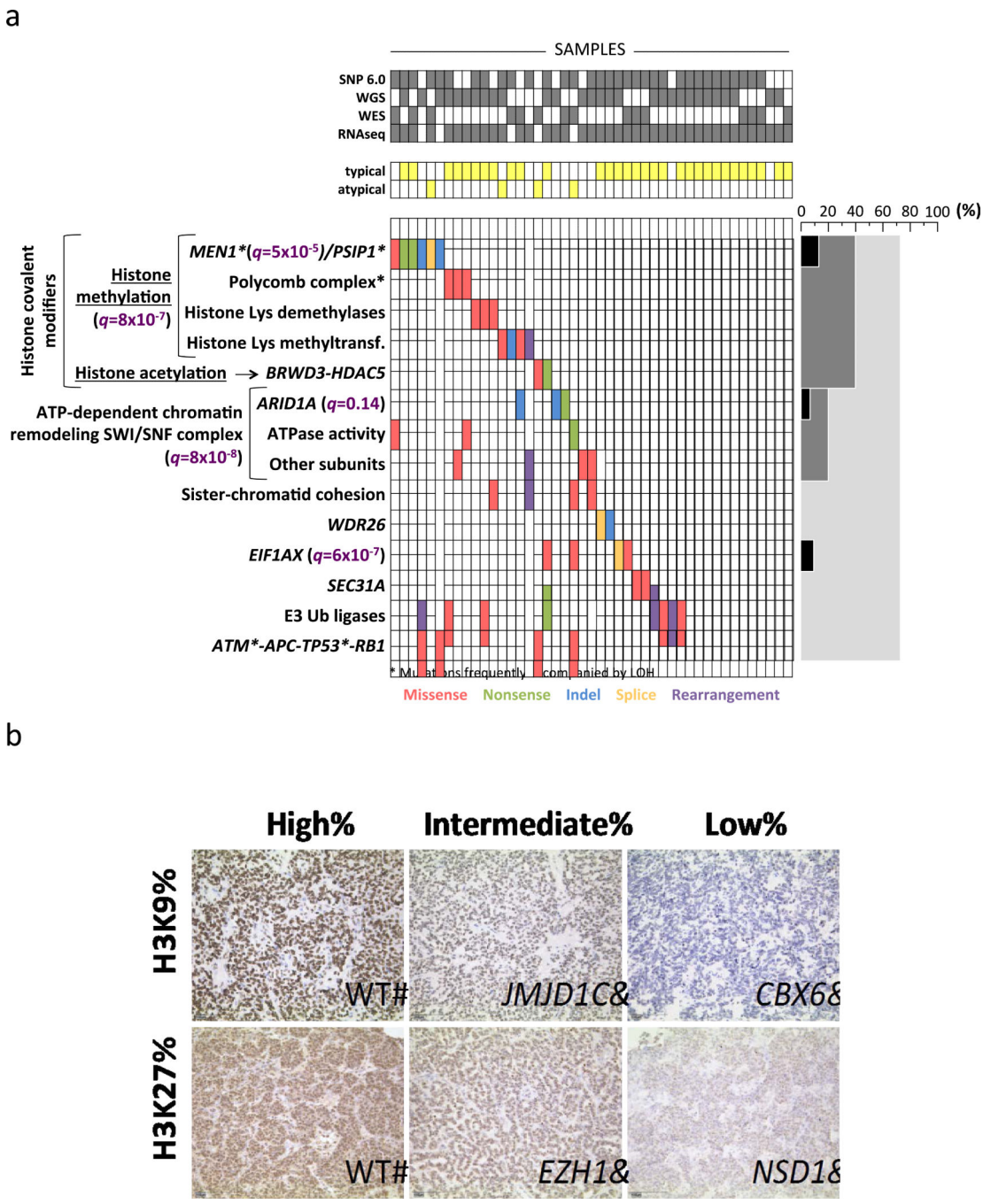


Figure 2.

Significant affected genes and pathways in pulmonary carcinoids. (a) Significantly mutated genes and pathways identified by genome (n=29), exome (n=15) and transcriptome (n=69) sequencing. The percentage of pulmonary carcinoids with a specific gene or pathway mutated is noted at the right side. The *q*-values of the significantly mutated genes and pathways are shown in brackets (see Methods section). Samples are displayed as columns and arranged to emphasize mutually exclusive mutations. (**b**) Methylation levels of H3K9me3 and H3K27me3 in pulmonary carcinoids. Representative pictures of different

degrees of methylation (high, intermediate, and low) for some of the samples summarized in Table 1. The mutated gene is shown in italics at the bottom right part of the correspondent picture. Wild-type samples are denoted by WT.

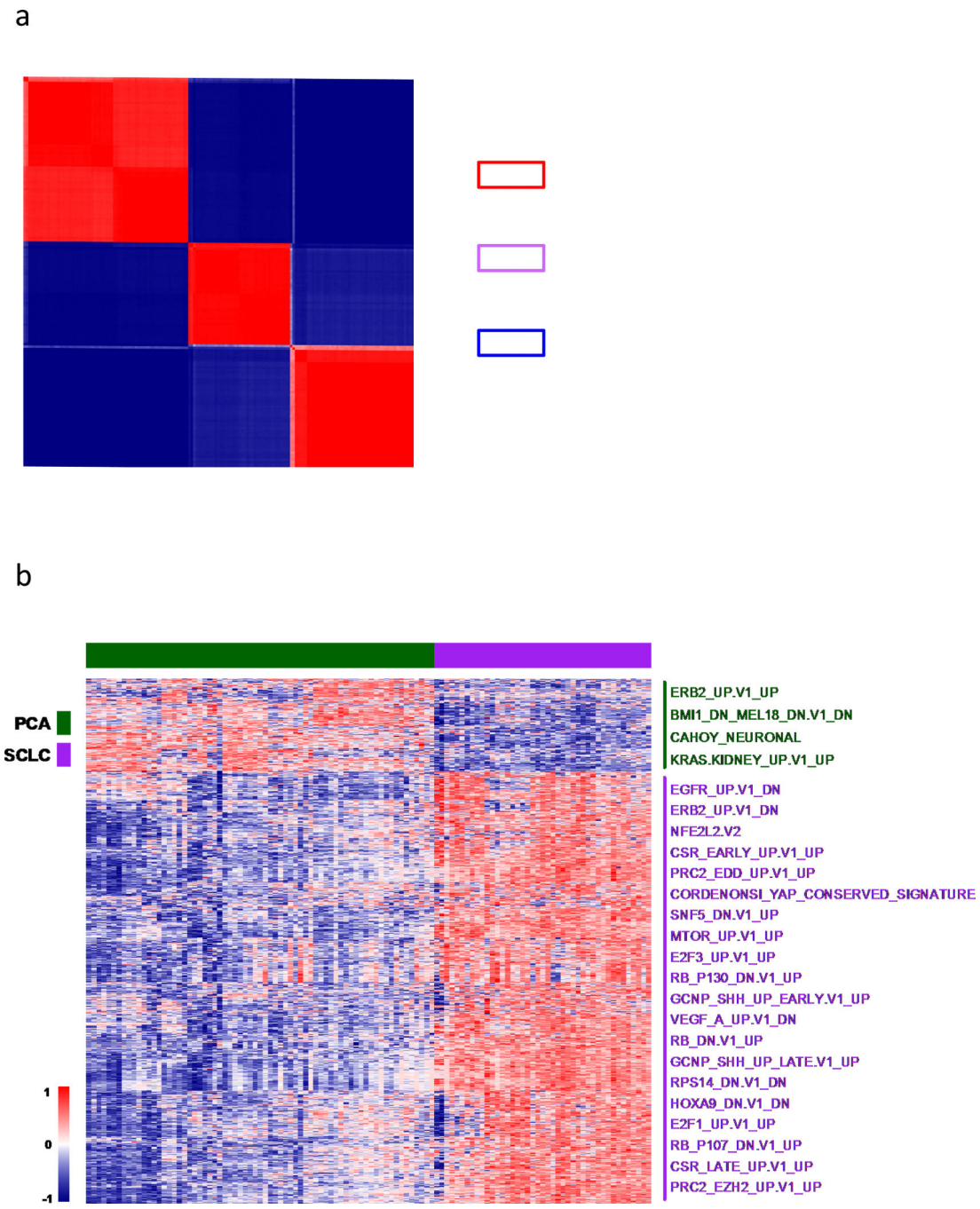


Figure 3.

Expression data analysis of pulmonary carcinois based on RNAseq data. (a) Consensus Kmeans clustering[32,33] using RNAseq expression data of 50 adenocarcinomas (AD, in blue), 42 small-cell lung cancer (SCLC, in red), and 69 pulmonary carcinoids (PCA, in purple) identified 3 groups using the clustering module from GenePattern[31] and consensus CDF[32,33] (left panel). The significance of the clustering was evaluated by using SigClust34 with a $p<0.0001$. Fisher's exact test35 was used to check associations between the clusters and the histological subtypes (right panel). (**b**) Gene Set Enrichment Analysis (GSEA)[21] for

SCLC versus PCA using RNAseq expression data. Low gene expression is indicated in blue and high expression, in red. On the right side are named the altered pathways in PCA (green) and SCLC (purple).

Table 1

Overview of samples annotated for mutations in genes involved in histone methylation, and correspondent levels of H3K9me3 and H3K27me3 detected by immunohistochemistry.

SAMPLE	MUTATION	H3K9me3	H3K27me3
S02333	*JMJD1C*_H954N	Intermediate	Low
S01502	*KDM4A*_I168T	Intermediate	N/A
S02323	*MEN1*_e3+1 and LOH	Low	Low
S02339	*NSD1*_A1047G	Intermediate	Low
S02327	*CBX6*_P302S and LOH	Low	Low
S01746	*EZH1*_R728G and LOH	Low	Intermediate
S02325	YY1_E253K	Low	Intermediate
S01501	Wild type	N/A	High
S01731	Wild type	Low	Low
S01742	Wild type	High	High
S02334	Wild type	Intermediate	High
S02337	Wild type	High	High
S02338	Wild type	High	Intermediate

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03099-x OPEN

Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors

Julie George et al.#

Pulmonary large-cell neuroendocrine carcinomas (LCNECs) have similarities with other lung cancers, but their precise relationship has remained unclear. Here we perform a comprehensive genomic ($n = 60$) and transcriptomic ($n = 69$) analysis of 75 LCNECs and identify two molecular subgroups: "type I LCNECs" with bi-allelic *TP53* and *STK11/KEAP1* alterations (37%), and "type II LCNECs" enriched for bi-allelic inactivation of *TP53* and *RB1* (42%). Despite sharing genomic alterations with adenocarcinomas and squamous cell carcinomas, no transcriptional relationship was found; instead LCNECs form distinct transcriptional subgroups with closest similarity to SCLC. While type I LCNECs and SCLCs exhibit a neuroendocrine profile with *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}, type II LCNECs bear *TP53* and *RB1* alterations and differ from most SCLC tumors with reduced neuroendocrine markers, a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high}, and an upregulation of immune-related pathways. In conclusion, LCNECs comprise two molecularly defined subgroups, and distinguishing them from SCLC may allow stratified targeted treatment of high-grade neuroendocrine lung tumors.

Correspondence and requests for materials should be addressed to J.G. (email: jgeorge@uni-koeln.de) or to E.B. (email: EBrambilla@chu-grenoble.fr) or to R.K.T. (email: roman.thomas@uni-koeln.de)
#A full list of authors and their affliations appears at the end of the paper

Molecular characterization studies have provided invaluable insight into the relationship between the major lung tumor subtypes[1–7]. These studies showed that morphologically defined lung adenocarcinomas, squamous cell carcinomas, and small cell carcinomas have distinct molecular phenotypes based upon their somatically altered genes[7]. Furthermore, global transcriptional analyses have revealed intra-group consistency, as well as substantial differences in the patterns of expressed genes, which led to the discovery of novel intra-group subtypes[2,3,8–11] and to the elimination of previous lung tumor categories (e.g., large-cell carcinoma)[7]. Of the remaining lung cancer subtypes, only large-cell neuroendocrine carcinomas (LCNECs) have so far not been characterized in depth using both transcriptomic, as well as genomic approaches.

LCNECs account for 2–3% of all resected lung cancers and belong to the category of neuroendocrine lung tumors, which also includes pulmonary carcinoids (PCa) and small cell lung cancer (SCLC)[12,13]. Contrary to pulmonary carcinoids, LCNEC and SCLC are clinically aggressive tumors presenting in elderly heavy-smokers with 5-year survival rates below 15–25% (LCNEC) and 5% (SCLC), respectively[12,13]. While therapy for both typical and atypical carcinoids and SCLC is primarily surgery and chemotherapy (in the case of SCLC), chemotherapy has limited efficacy in LCNEC patients and no standard treatment regimen exists for this tumor type[14]. Thus, LCNECs share both commonalities (e.g., neuroendocrine differentiation) and discrepancies (e.g., limited response to chemotherapy) with SCLC; however, the underlying molecular basis of these shared and distinct features is only poorly understood. Further complicating the histological classification, LCNECs are sometimes found combined with adenocarcinoma or squamous cell carcinoma and some SCLCs are combined with a component of LCNEC[12,13]. Thus, defining the molecular patterns of this tumor type presents the opportunity to not only reveal possible novel therapeutic targets, but also help clarifying the ontogeny and relationship of lung tumors in general.

Previous efforts in characterizing LCNECs through targeted sequencing of selected cancer-related genes[15–17] and through gene expression profiling[18] provided some first insights; however, global genomic studies combined with transcriptomic analyses have so far been lacking. Furthermore, given the lack of adequate therapeutic strategies in LCNECs, a precise delineation of the molecular boundaries between different neuroendocrine tumors is needed. We therefore aimed to comprehensively dissect both the mutational and the transcriptional patterns of this tumor type.

In this report, we show that LCNECs are composed of two mutually exclusive subgroups, which we categorize as "type I LCNECs" (with *STK11*/*KEAP1* alterations) and "type II LCNECs" (with *RB1* alterations). Despite sharing genomic alterations with lung adenocarcinomas and squamous cell carcinomas, type I LCNECs exhibit a neuroendocrine profile with closest similarity to SCLC tumors. While type II LCNECs reveal genetic resemblance to SCLC, these tumors are markedly different from SCLC with reduced levels of neuroendocrine markers and high activity of the *NOTCH* pathway. Conclusively, LCNECs represent a distinct subgroup within the spectrum of high-grade neuroendocrine tumors of the lung, and our findings emphasize the importance of distinguishing LCNECs from other lung cancers subtypes.

Results

Genomic alterations in LCNECs. We collected 75 fresh-frozen tumor specimens from patients diagnosed with LCNEC under institutional review board approval (Supplementary Data 1). All tumors were thoroughly analyzed, and the histological features of pulmonary LCNECs were confirmed by expert pathologists (E.B., W.T., R.B.) according to the 2015 WHO classification[13] (Supplementary Data 2). Most tumors were obtained from current or former heavy smokers, and enriched for stages I and II (68%). Nineteen of 75 LCNECs included in this study showed additional histological components of lung adenocarcinoma (ADC) ($n = 2$), squamous cell carcinoma (SqCC) ($n = 5$) or SCLC ($n = 12$) (Supplementary Data 1–2). In subsequent analyses nucleic acids were extracted only from pure LCNEC regions (Methods section).

Early genomic profiling studies employing targeted sequencing of selected cancer-related genes aided in the identification of some prevalent mutations in LCNECs[15–17]. In order to assess globally all genomic alterations in LCNECs and to compare them to those occurring in other lung tumors, we conducted whole-exome sequencing (WES) of 55 LCNEC tumor-normal pairs; we additionally performed whole-genome sequencing (WGS) in those cases where sufficient material was available ($n = 11$), thus amounting to sequencing data of 60 LCNECs in total (six tumors were both, genome- and exome-sequenced, Supplementary Fig. 1a). We furthermore performed Affymetrix 6.0 SNP array analyses of 60 and transcriptome sequencing of 69 tumors (Supplementary Data 1; Supplementary Fig. 1a). Despite initial review to include cases with a microscopic tumor content of >70%, sequencing data analysis revealed a median tumor purity of 59.5% and a median ploidy of 2.8 (Supplementary Data 1, Supplementary Fig. 1b, Methods section). On average, LCNECs exhibited an exonic mutation rate of 8.6 non-synonymous mutations per million base pairs and a C:G > A:T transversion rate of 38.7% (Fig. 1a, Supplementary Data 1), indicative of tobacco exposure[1–6]. We analyzed the signatures of mutational processes[19,20] in LCNECs, which confirmed a prominent smoking-related signature (signature 4[19,20]) that accounts for the majority of all somatic mutations, and which is in general comparable to most other lung tumors of heavy smokers (Supplementary Fig. 1c–f, Supplementary Data 3).

Analyses of chromosomal gene copy numbers revealed statistically significant amplifications of 1p34 (containing the *MYCL1* gene, 12%), 8p12 (containing *FGFR1*, 7%), 8q24.21 (containing *MYC*, 5%), 13q33 (containing *IRS2*, 3%), and 14q13 (containing *NKX2-1*, also known as TTF-1, 10%) ($Q < 0.01$, Supplementary Fig. 2a; Supplementary Data 4–5, Methods section). Statistically significant deletions affected *CDKN2A* (9p21, 8%) and a putative fragile site at *PTPRD* (9p24, 7%)[21]. While amplifications of *NKX2-1* and *FGFR1* frequently occur in lung adenocarcinomas[1,2,7,21] and squamous cell carcinomas[3,7,21,22], respectively, *MYCL1* amplifications are commonly found in SCLC[4–6,23]. Thus, LCNECs harbor significant copy-number alterations that occur in different lung cancer subtypes.

We next applied analytical filters to identify mutations with biological relevance in the context of a high-mutation rate and found eight significantly mutated genes ($Q < 0.01$, Methods section, Fig. 1a, Supplementary Data 6–7). *TP53* was the most frequently mutated gene (92%), followed by inactivating somatic events in *RB1* (42%); bi-allelic alterations in both genes, *TP53* and *RB1*—a hallmark of SCLC[4–6]—were found in 40% of the cases (Supplementary Fig. 2b, Supplementary Data 6–9). Notably, LCNECs with admixtures of other histological components mostly had *RB1* alterations (Fig. 1a). While genomic alterations in *RB1* resulted in loss-of-nuclear Rb1 expression ($P < 0.0001$, Fisher's exact test, Supplementary Fig. 3a), immunohistochemistry revealed that absence of Rb1 was not only confined to the LCNEC component, but also evident in the combined other histological subtype (6/7 cases, Supplementary Fig. 3b, Supplementary Data 2). This may implicate shared genetic features between LCNECs and the admixtures of other histological carcinoma types.

We furthermore identified—frequently deleterious—somatic alterations in functionally relevant domains of *STK11* (30%) and

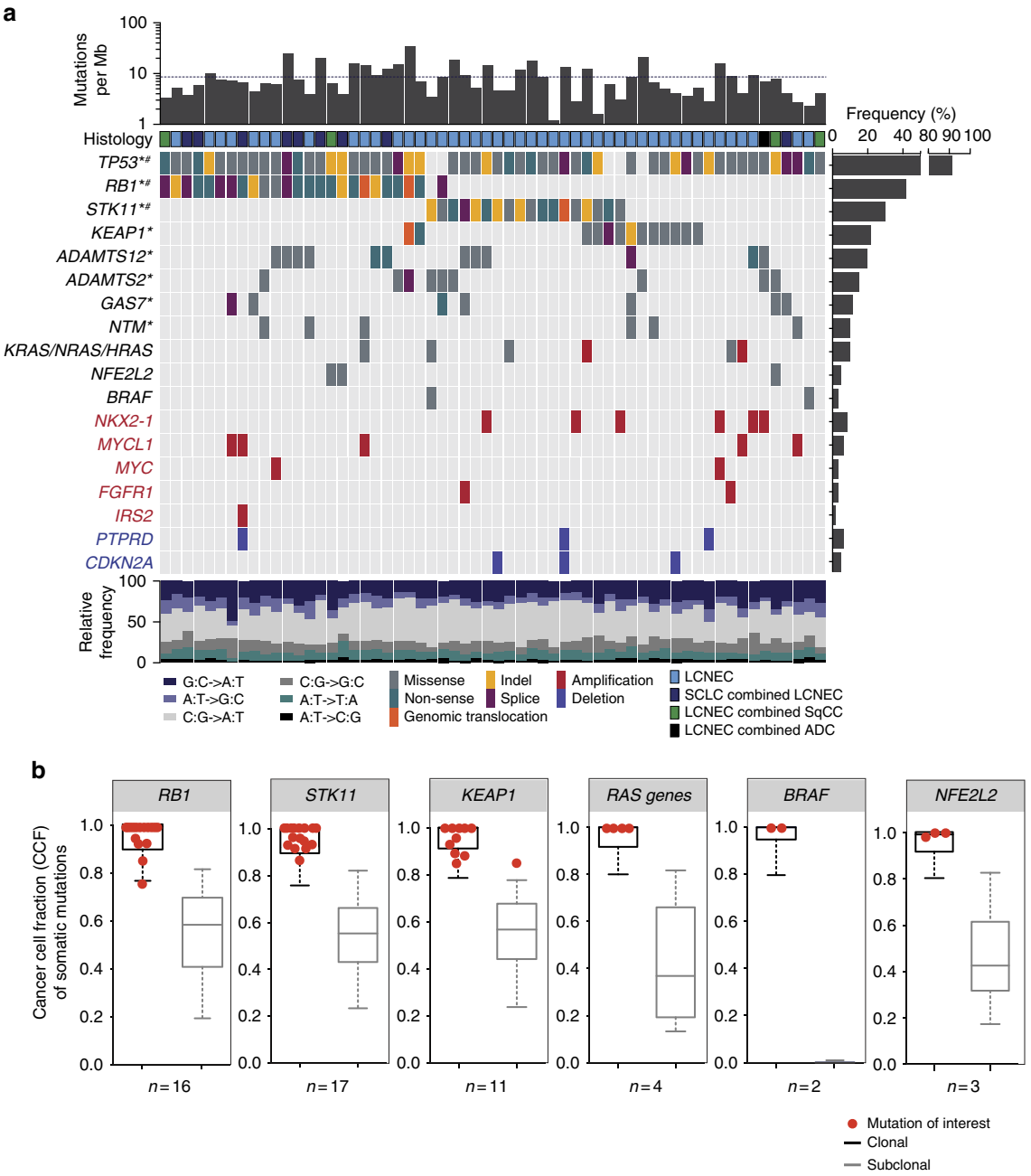
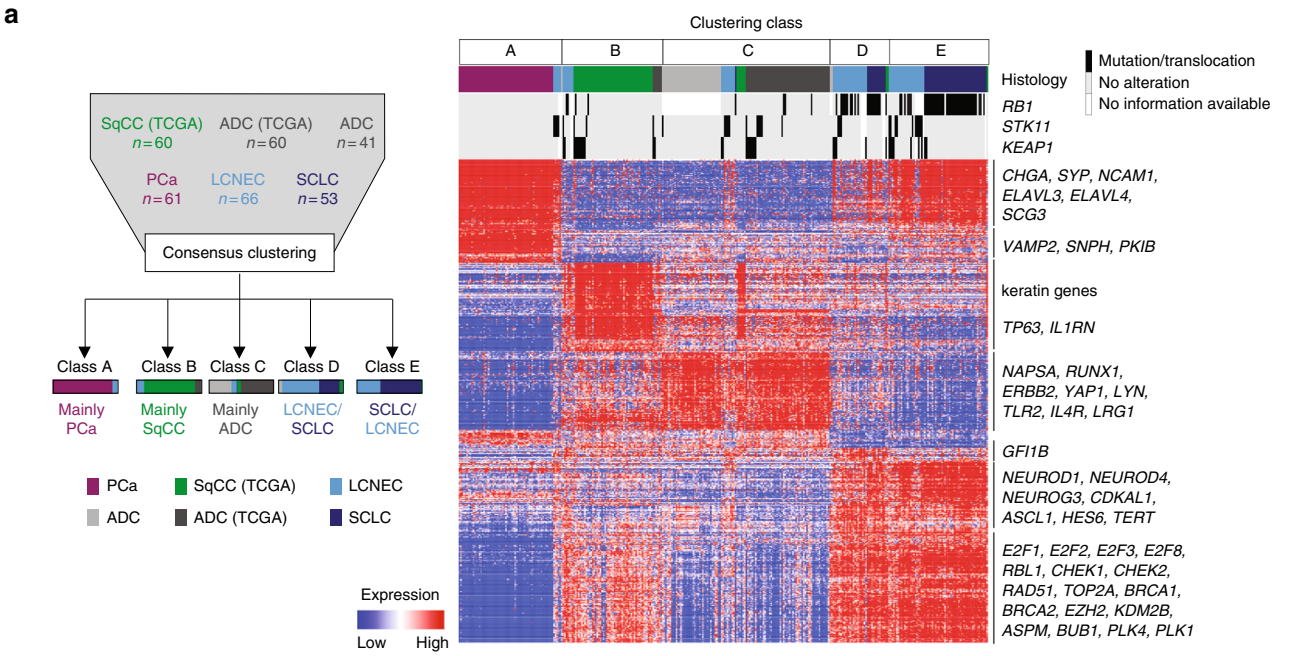


Fig. 1 Genomic alterations in pulmonary large-cell neuroendocrine carcinomas (LCNECs). **a** Tumor samples are arranged from left to right. Histological assignments and somatic alterations in candidate genes are annotated for each sample according to the color panel below the image. The somatic mutation frequencies for each candidate gene are plotted on the right panel. Mutation rates and the type of base-pair substitutions are displayed in the top and bottom panel, respectively; a dashed black line indicates the average value. Significantly mutated genes and genes with a significant enrichment of damaging mutations are denoted with * and #, respectively ($Q < 0.01$, Methods section). Genes with significant copy number (CN) amplifications (CN > 4) and deletions (CN < 1) (Supplementary Fig. 2a, Supplementary Dataset 5) are displayed in red and blue, respectively ($Q < 0.01$, Methods section). **b** The distribution of clonal and sub-clonal mutations was analyzed for tumor samples that harbored mutations in key candidate genes. The cancer cell fractions (CCF) of all mutations were determined, assigned to clonal or sub-clonal fractions (Methods section), and displayed as whiskers box-plot (median and interquartile range, whiskers: 5–95 percentile). The CCF of candidate gene mutations is highlighted in red

KEAP1 (22%)[1–3] (Fig. 1a, Supplementary Fig. 4a, Supplementary Data 6–9). Combined with loss-of-heterozygosity (LOH), bi-allelic alterations of *STK11* and *KEAP1* were found in 37% of the cases (Supplementary Fig. 2b, Supplementary Data 8). In those cases where WGS was performed, we were able to identify larger genomic rearrangements, which led to the inactivation of *RB1*, *STK11*, or *KEAP1* (Fig. 1a, Supplementary Fig. 4a, Supplementary Data 9). Altogether, somatic alterations of *RB1* and *STK11*/*KEAP1* were detected in 82% of the cases ($n = 49$) and occurred in a mutually exclusive fashion ($P < 0.0001$, Fisher's exact test, Fig. 1a). We furthermore observed a trend toward inferior outcome in patients with *RB1*-mutated tumors ($P = 0.126$, log-rank test, Supplementary Fig. 4b). The genomic profiling thus points to two distinct subgroups of LCNECs.

We additionally identified statistically significant mutations in the metalloproteinases *ADAMTS2* (15%) and *ADAMTS12* (20%),



b

Biological processes and pathways

Class A (mainly PCa)	↑ Transmission of nerve impulse, synaptic transmission ↑ Cellular respiration, generation of metabolites and energy	↓ Immune response, inflammatory response ↓ Cell cycle process, cell cycle
Class B (mainly SqCC)	↑ Keratinocyte and epidermal cell differentiation ↑ Mitotic and cell cycle process ↑ Immune response	↓ Transmission of nerve impulse, ↓ Neuron differentiation and development
Class C (mainly ADC)	↑ Immune and inflammatory response, chemotaxis ↑ Cell adhesion	↓ Cell cycle process and M phase ↓ Response to DNA damage, DNA repair
Class D (LCNEC/SCLC)	↑ Cell cycle, mitosis ↑ Response to DNA damage, DNA repair ↑ Chromatin modification	↓ Generation of metabolites and energy ↓ Oxidation reduction and phosphorylation
Class E (SCLC/LCNEC)	↑ Cell cycle, mitosis ↑ Chromatin organization	↓ Immune and inflammatory response, chemotaxis ↓ Cell adhesion

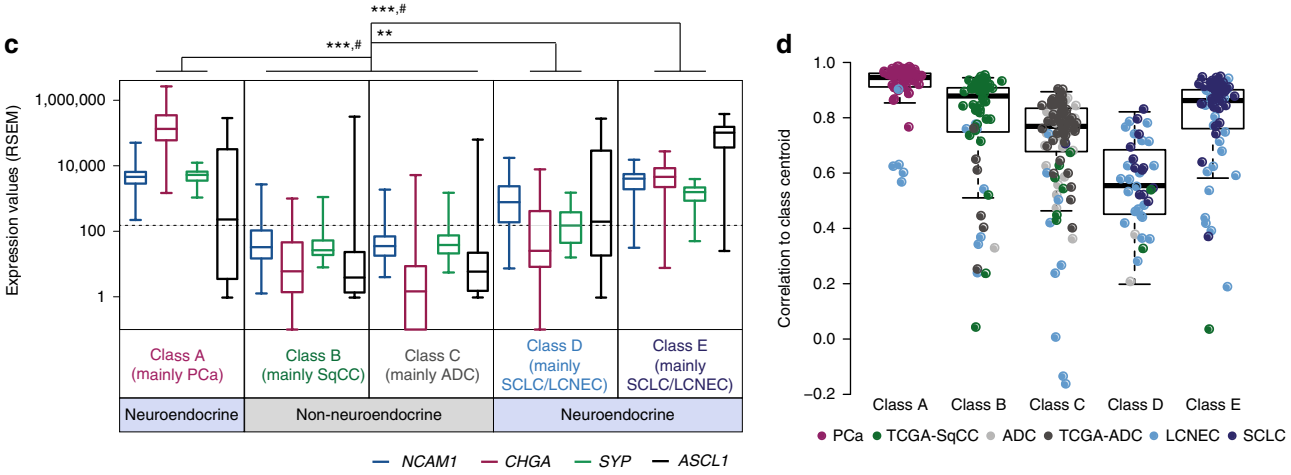


Fig. 2 Gene expression studies on lung cancer subtypes. **a** A schematic description of the unsupervised consensus clustering approach is provided on the left panel. The clustering results are displayed on the right panel as a heatmap, in which tumor samples are arranged in columns, grouped according to their expression clustering class, annotated for the histological subtype and for the somatic alteration status. Expression values of genes identified by ClaNC (Methods section) are represented as a heatmap; red and blue indicate high and low expression, respectively. Selected candidate genes are shown on the right. **b** Significant enrichment of differentially expressed genes in signaling pathways is displayed for all clustering classes ($P < 0.0001$, Methods section). **c** Expression values for key neuroendocrine differentiation markers are plotted for each clustering class as box-plots (median and interquartile range, whiskers: min–max values). Dashed black lines indicate the threshold for low expression (Methods section). $Q < 0.05$ (#), significance determined by SAM (Supplementary Dataset 12); $P < 0.001$ (***) Mann–Whitney *U*-test. **d** The correlation of each sample to the centroid of its clustering class was calculated and displayed as box-plot (median and interquartile range, whiskers 5–95 percentile)

and in *GAS7* (12%) and *NTM* (10%) ($Q < 0.01$, Methods section, Fig. 1a, Supplementary Fig. 4c, Supplementary Data 6–7), which so far have not been reported as significantly mutated in any other lung cancer subtype. The mutations affected functionally important protein domains, which may suggest a relevant role in the tumorigenesis of LCNECs (Supplementary Fig. 3c).

We also analyzed LCNECs for alterations in genes of known tumor-specific functions (e.g., *CREBBP*, *EP300*[3,4,6,21], *NOTCH*[3,6,21], *MEN1*[24], *ARID1A*[1–3,21,24]) (Supplementary Fig. 2b, Supplementary Fig. 4d, Supplementary Data 6) and found oncogenic mutations of *RAS* family genes (*KRAS*-G12V, -G12C, *NRAS*-D57E, *HRAS*-G13R), *NFE2L2* (2 cases with G31V and 1 case with E79Q) and *BRAF* (V600E, and G469V). Combined with focal amplifications, *RAS* genes were affected in 10% of the tumors (Fig. 1a; Supplementary Data 5–6). We also identified several private in-frame fusion events, e.g., involving the kinases *NTRK1* and *PTK6*, which were, however, not recurrent (Supplementary Fig. 5, Supplementary Data 10). Thus, LCNECs harbor alterations of oncogenes which are commonly found in lung adenocarcinomas, but usually absent in neuroendocrine tumors like SCLC.

The distinct mutational patterns in LCNECs and the presence of other histological components may suggest a high level of intra-tumor heterogeneity. We analyzed the clonal distribution of somatic alterations and determined the cancer cell fraction (CCF) of each somatic mutation call (Methods section). Despite the fact that some LCNECs were found with admixtures of other histological subtypes (Fig. 1a, Supplementary Data 1–2), our studies on the LCNEC component of such composite tumors pointed to little intra-tumor heterogeneity with a median of 7% sub-clonal mutations per sample (Supplementary Fig. 2b–c, Supplementary Data 1, Methods section). Furthermore, all relevant and significant mutations were found to be clonal within the tumor, thus suggesting these alterations as early events during tumorigenesis (Fig. 1b, Supplementary Data 6).

In summary, genome sequencing revealed distinct genomic profiles in LCNECs. While certain alterations (e.g., *RB1, MYCL1*) resemble patterns found in SCLC[4–6,23], others are typical of lung adenocarcinoma or squamous cell carcinomas (e.g., *STK11*, *KEAP1*, *NKX2-1*, *RAS*, *BRAF*, and *NFE2L2*)[1–3,7,21]. Thus, LCNECs appear to divide into molecularly defined subsets of tumors with genomic similarities to other major lung cancer subtypes.

Transcriptional profiles of LCNECs and other lung cancers. Our sequencing efforts have revealed genomic alterations in LCNECs that were previously known as canonical alterations in either, lung adenocarcinomas, squamous cell carcinomas[7,21], or SCLC[4–6]. In light of these distinct associations, it remained to be understood if these genomic correlates might reflect a relationship of LCNECs with these lung tumor subtypes on the level of gene expression. We therefore analyzed whether the transcriptional patterns in LCNECs are correlated with the expression profiles of other lung cancers.

We compared the expression data of LCNECs with lung adenocarcinomas[2,3,25–27], squamous cell carcinomas[3], SCLC[6] and pulmonary carcinoids[24] following extensive normalization of the transcriptome sequencing data (Fig. 2a, Methods section, Supplementary Data 11). Unsupervised consensus clustering yielded five consistent expression clusters, which correlated with the histological annotation of the tumors ($P < 0.0001$, Fig. 2a, Supplementary Fig. 6–7, Supplementary Data 12): pulmonary carcinoids, squamous cell carcinomas and adenocarcinomas formed distinct transcriptional classes (classes A, B, and C, respectively), with few LCNECs falling into these groups. However, the majority of SCLC and LCNECs clustered in two transcriptional subgroups (classes D and E) (Fig. 2a); a phenomenon that had previously been observed in other studies on high-grade neuroendocrine tumors[6,18]. While the majority of SCLC tumors formed consensus cluster E (75% of all SCLC cases analyzed), a fraction of SCLC tumors shared transcriptional similarities with LCNECs that predominantly formed cluster D. Thus, LCNECs appear to be more closely related to SCLCs than to adenocarcinomas or squamous cell carcinomas.

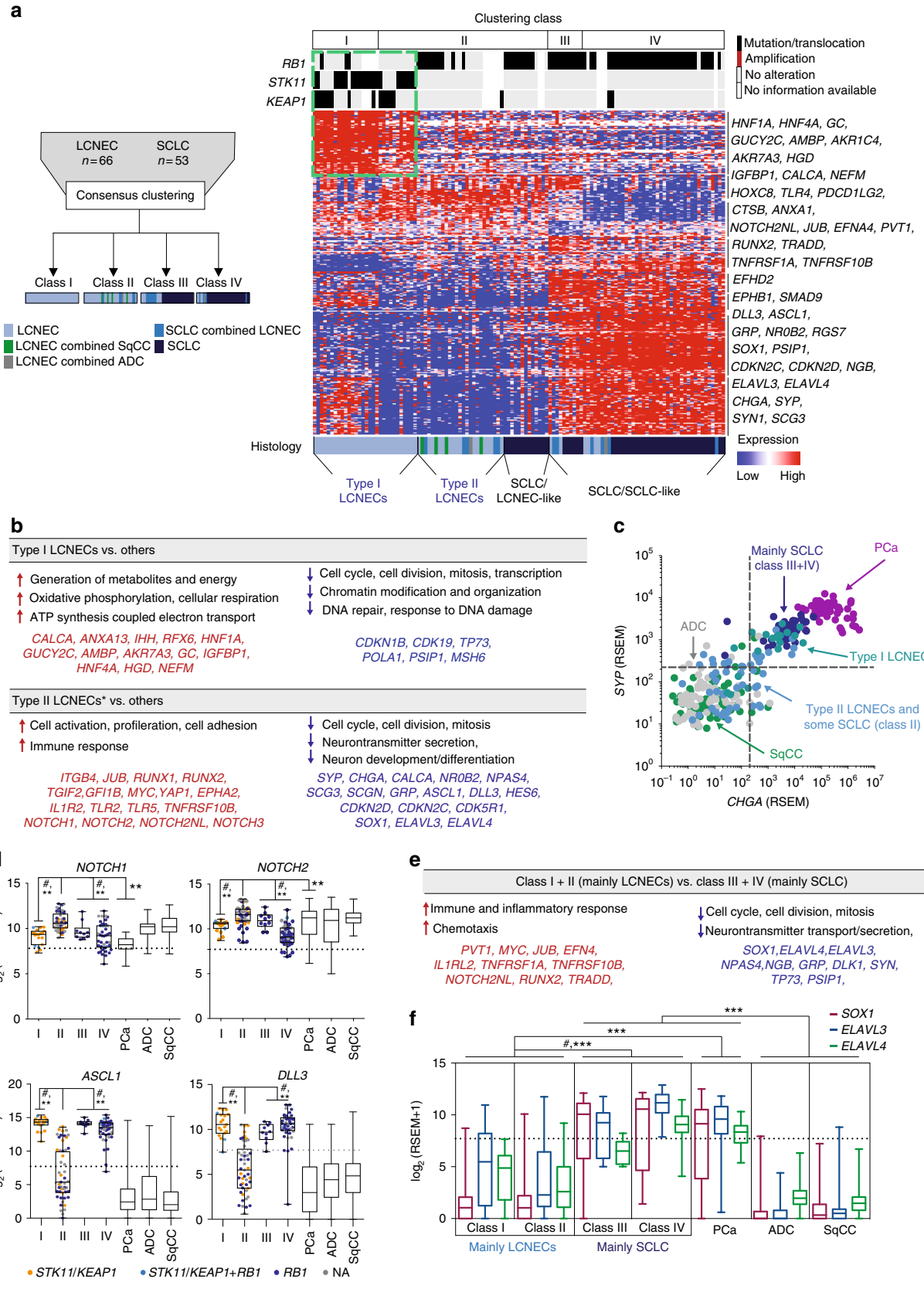
We next analyzed the transcriptome sequencing data for differentially expressed genes and their enrichment in biological pathways (Methods section). In line with previous observations[2,3,9–11,18,28], this analysis showed that both adenocarcinomas and squamous cell carcinomas exhibited upregulation of pathways controlling cell differentiation, adhesion and immune responses, along with higher expression of *ERBB2* and *TP63* (Fig. 2b, Supplementary Fig. 8a, Supplementary Data 13–14, $Q < 0.05$, Methods section). Lung neuroendocrine tumors, on the contrary, showed significantly higher expression of neuroendocrine and endocrine markers, Hu antigens (*ELAVL3* and *ELAVL4*) and the lineage transcription factor and oncogene *ASCL1*, which is in agreement with previous studies on lung cancer subtypes[11–13,18,29] ($Q < 0.05$, Methods section). Furthermore, particularly high expression of the neuronal and endocrine lineage transcription factors *NEUROD1*, *NEUROD4*, and *NEUROG3*[30,31] was found in SCLC and LCNECs of transcriptional class E (Fig. 2a, c, Supplementary Fig. 8b–e, Supplementary Data 13, $Q < 0.05$). While recent studies employing SCLC cell lines and mouse models indicated discordant expression patterns for *ASCL1* and *NEUROD1*[31], our sequencing data of human high-grade neuroendocrine lung tumors revealed expression of both neuroendocrine lineage factors in class E (Supplementary Fig. 8f).

Within the spectrum of neuroendocrine lung tumors, pulmonary carcinoids formed a distinct subgroup with functional enrichment in pathways regulating cellular respiration and metabolism. LCNECs mostly shared similarities with SCLC, revealing upregulation of pathways and genes controlling cell cycle and mitosis (E2F transcription factors and checkpoint kinases), DNA damage response (*RAD51*, *TOP2A*, and *BRCA1*) and centrosomal functions (such as *BUB1*, *PLK1*, and *ASPM*); which, to some extent, were also found in squamous cell carcinomas (Fig. 2b; Supplementary Fig. 8g–i, Supplementary Data 13–14), and which is in agreement with previous studies[18]. Further supporting a molecular relationship of SCLC and LCNECs in a fraction of the cases, *RB1*-mutated LCNECs were enriched in classes D and E ($P < 0.05$, Fisher's exact test). Although, LCNECs also harbored alterations commonly observed in adenocarcinomas and squamous cell carcinomas, even LCNECs with such alterations in *KEAP1* or *STK11* were primarily found in transcriptional subclasses shared with SCLC (Fig. 2a, Supplementary Fig. 7c, Supplementary Data 12). Therefore, this observation supports the view that despite the similarity in oncogenic mutations, LCNECs rather constitute their own biological class; and may not be considered as neuroendocrine versions of adenocarcinomas or squamous cell carcinomas.

We also quantified the consistency of the expression profiles for each sample with respect to its clustering group. Again, this analysis revealed a strong correlation for most LCNECs clustering with SCLC tumors (classes D and E); on the other hand, expression profiles of those few LCNEC samples clustering with lung adenocarcinomas, squamous cell carcinomas, and pulmonary carcinoids were less consistent (Fig. 2d). Furthermore, we performed separate transcriptional clustering of LCNECs with adenocarcinomas and squamous cell carcinomas only (excluding SCLC), which did not suggest any unrecognized similarities between these lung cancer subtypes (Supplementary Fig. 9). Thus,

despite sharing somatic alterations with other tumor subtypes, such as adenocarcinomas and squamous cell carcinomas, LCNECs were transcriptionally dissimilar with all non-neuroendocrine lung tumors and showed closest similarities to SCLC.

The transcriptional relationship of LCNEC and SCLC. In the previous section, we sought for a global approach to identify common and distinct transcriptional profiles of LCNECs in relationship with other lung tumors, which showed that LCNEC and SCLC appear to share most transcriptional patterns.



However, strongly divergent tumors (e.g., carcinoids, adenocarcinomas) may drive these clusters and mask important differences between LCNECs and SCLC. We therefore sought to directly compare LCNECs and SCLC on the transcriptional level (Fig. 3a). The resulting unsupervised clustering analysis revealed four consensus clusters of LCNEC and SCLC that we termed classes I–IV in order to distinguish them from the above-mentioned classes A–E (Fig. 3a, Supplementary Fig. 10–11, Supplementary Data 12). Class I exclusively included LCNECs with *STK11* or *KEAP1* alterations; yet, a few cases with these alterations fell into class II that predominantly consisted of LCNECs with *RB1* loss (Fig. 3a). Some LCNECs, including tumors admixed with SCLC ("SCLC combined LCNECs")—clustered with the majority of SCLC tumors in the classes III and IV; similarly, some SCLC tumors were part of class II that included LCNECs bearing *RB1* alterations (Fig. 3a, Supplementary Fig. 11). Even though pathological review had been conducted to distinguish histological subtypes from one another, transcriptional clustering suggested high degrees of similarity for some LCNEC and SCLC cases; these tumors may therefore be considered as "SCLC-like" and "LCNEC-like" (Fig. 3a, Supplementary Fig. 11, Supplementary Data 11). Other major genome alterations (e.g., *NKX2-1*, *MYCL1*, *RAS* genes, *NFE2L2*, *BRAF*) did not segregate with the identified transcriptional subgroups (Supplementary Fig. 11). We further analyzed the consistency of the transcriptional subgroups by clustering LCNECs alone, which revealed high concordance with the transcriptional classes identified in Fig. 3a (62/66 cases, 94%, $P < 0.001$, Fisher's exact test, Supplementary Fig. 13, Supplementary Data 12). Thus, despite the similarities between LCNECs and SCLCs, subtypes of LCNECs exist with profound differences to SCLC.

The transcriptional clustering heatmap pointed to a strong gene expression pattern shared by all LCNECs bearing *STK11/KEAP1* alterations (Fig. 3a, Supplementary Fig. 12a, green box in upper left quadrant). We therefore conducted a supervised analysis of the gene expression data, in which LCNECs with *STK11/KEAP1* alterations were compared to tumors bearing *RB1* alterations. This analysis indicated specific expression profiles, which were similar to those observed in tumors constituting class I (Fig. 3b, Supplementary Fig. 12, Supplementary Data 13). We therefore assigned this genomic subset of tumors to one group, termed "type I LCNECs".

Type I LCNECs exhibited high levels of calcitonin A (*CALCA*), a known marker of pulmonary neuroendocrine cells[32–34] (Fig. 3a, Supplementary Fig. 12b, Supplementary Data 13). This subgroup furthermore displayed a pronounced upregulation of cellular metabolic pathways, which we also observed in pulmonary carcinoids (Fig. 2b), but which was less prominent in LCNECs and SCLC tumors with *RB1* alterations (Fig. 3a, b, Supplementary Data 12–13). Other genes found in type I LCNECs included gastrointestinal transcription factors (e.g., *HNF4A*, *HNF1A*, and *RFX6*), which were previously reported to play a role in de-differentiated lung tumors[35,36] (Fig. 3b, Supplementary Fig. 12c, d, Supplementary Data 13).

The most striking difference was found in the expression levels of neuroendocrine genes: while type I LCNECs and the majority of SCLC tumors (class III + IV) harbored high levels of neuroendocrine genes (*CHGA* and *SYP*; Fig. 3c; Supplementary Fig. 12e; Supplementary Data 12), most LCNECs and some SCLC tumors with *RB1* alterations in class II exhibited low levels of these genes (Fig. 3c, Supplementary Fig. 12e). By contrast, tumors in class II displayed elevated expression of genes associated with active Notch signaling (e.g., *NOTCH1*, *NOTCH2*, and *HES1*) and immune cell responses (e.g. *PDCD1LG2*, *TLR4*, and *CTSB*) (Fig. 3a, d, Supplementary Fig. 12f, Supplementary Data 12–13). Given the strong enrichment of LCNECs with *STK11* or *KEAP1* alterations in cluster I, and the prominent lack of expression of key neuroendocrine genes in most tumors of class II, we termed LCNECs within this transcriptional class as "type II LCNECs".

We have recently demonstrated that SCLC tumors usually harbor inactive Notch signaling and that activation of Notch reduces expression of neuroendocrine genes (e.g., *CHGA*, *SYP* and *NCAM1*) and Ascl1[6]. Consistent with this notion, we found that type II LCNECs and some SCLC within this transcriptional class exhibited signs of *NOTCH* upregulation and low expression of neuroendocrine markers, *ASCL1* and *DLL3*, an inhibitor of the Notch signaling pathway[37] (Fig. 3d, and Supplementary Fig. 12f). Conversely, type I LCNECs and the majority of the SCLC samples (class III and IV) showed higher levels of neuroendocrine genes, as well as of *ASCL1* and *DLL3*, and downregulation of *NOTCH* pathway genes (Fig. 3d, Supplementary Fig. 12f). Thus, despite the fact that type II LCNECs and some SCLCs harbor bi-allelic loss of *TP53* and *RB1*, their transcriptional signatures include low levels of neuroendocrine genes and a distinct profile of $NOTCH^{high}$ and $ASCL1^{low}/DLL3^{low}$, which differentiates these tumors from type I LCNECs and from the majority of SCLC cases. We did not identify any significant enrichment of somatic alterations in *NOTCH* pathway genes, which may explain these transcriptional differences (Supplementary Fig. 11). However, a recent study in a pre-clinical mouse model has established a central role of *REST* as a repressor of neuroendocrine markers in SCLC[38]. Compatible with these findings, type II LCNECs displayed significantly higher levels of *REST* (clustering class II, Supplementary Data 12, $Q < 0.05$), which may explain the low neuroendocrine phenotype in type II LCNECs marked by $ASCL1^{low}/DLL3^{low}/NOTCH^{high}$. Given the important role of NOTCH signaling and *ASCL1* in the decision of neuroendocrine fate and the development of neuroendocrine lung tumors[29,31,38], these findings provide further support for our distinction of type I and II LCNECs.

We next analyzed the relationship of the expression classes I–IV using hierarchical clustering, which revealed two major subgroups (Supplementary Fig. 11): one subgroup mainly consisting of LCNECs (type I and II LCNECs), and the other subgroup mainly containing SCLC tumors (classes III and IV). Thus, despite harboring distinct transcriptional subcategories, LCNEC and SCLC tumors largely followed their histological annotation and formed separate transcriptional subgroups. Differentially expressed genes included *SOX1* and the neuroendocrine Hu genes (*ELAVL3*,

Fig. 3 Gene expression studies on LCNEC and SCLC. **a** The expression profiles of LCNEC and SCLC tumors were analyzed following the annotation and approach described in Fig. 2a. Expression values of genes identified by ClaNC (Methods section) are represented as a heatmap in which red and blue indicate high and low expression, respectively. Selected candidate genes are shown on the right. Dashed green lines indicate an expression profile shared by LCNEC tumors with *STK11/KEAP1* alterations (type I LCNECs). **b** The significant enrichment of differentially expressed genes and signaling pathways are displayed for type I LCNECs and type II LCNECs. $P < 0.0001$ (Methods section); * some SCLC tumors that co-clustered with type II LCNECs were included in this analysis. Key candidate genes are highlighted in bold. **c**, **d** Expression values for **c** the key neuroendocrine differentiation markers *SYP* (synaptophysin) and *CHGA* (chromogranin A) (scatter plot), and **d** *NOTCH* pathways genes (box plots: median and interquartile range, whiskers: min–max values). **e** Significant enrichment of differentially expressed genes and signaling pathways was analyzed for class I and II vs class III and IV tumor samples; $P < 0.0001$ (Methods section). **f** Expression values of *SOX1*, *ELAVL3*, and *ELAVL4* are plotted for the clustering classes and other lung cancer subtypes (box plots: median and interquartile range, whiskers: min–max values). $Q < 0.05$ (#), SAM (Supplementary Dataset 12); $P < 0.01$ (**) Mann–Whitney *U*-test

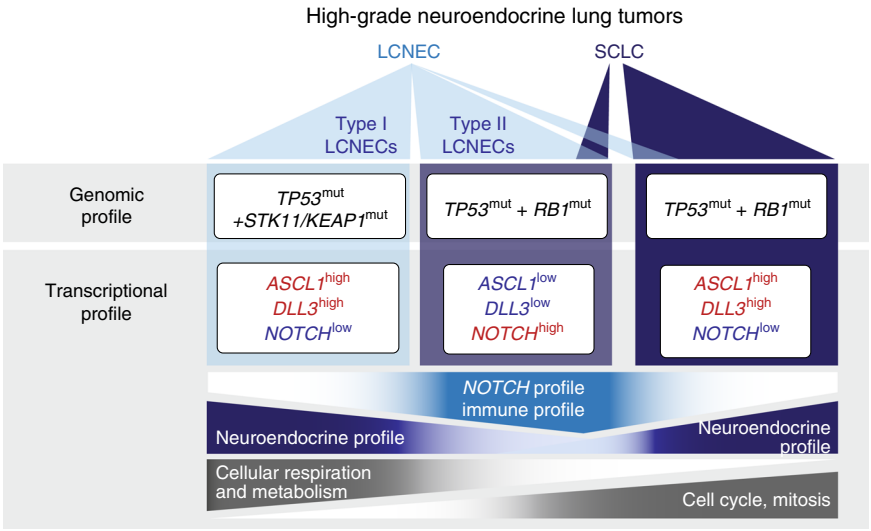


Fig. 4 Schematic overview of somatic alterations and expression profiles in high-grade neuroendocrine lung tumors. Significantly mutated genes are shown in black and differentially expressed genes are highlighted in red and blue, describing higher and lower expression, respectively. Upregulated expression profiles and signaling pathways are indicated by color gradients

ELAVL4), which were enriched in most SCLC samples (classes III and IV (Supplementary Data 13, $Q < 0.05$, Methods section) (Fig. 3f). This observation is in line with previous reports on autoantibodies against Sox1 and Hu-proteins that are commonly found in SCLC patients[39]. While pulmonary carcinoids harbored similar expression levels, these genes were essentially absent or only moderately expressed in most LCNECs and other lung cancer subtypes (Fig. 3f).

We furthermore analyzed the impact of transcriptional subgroups on tumor stage and clinical outcome. While, we found no association of tumor stage with the molecular subsets found in high-grade neuroendocrine tumors (Supplementary Data 12), we observed a trend toward inferior survival in patients with SCLC (transcriptional profiles of classes III and IV; $P = 0.072$, log-rank test, Supplementary Fig. 14), which was similarly observed in previous studies on high-grade neuroendocrine lung tumors[18].

Conclusively, LCNECs exhibit a distinct expression profile within the spectrum of high-grade neuroendocrine lung tumors, which can further be divided into two subtypes: type I LCNECs with high neuroendocrine expression and, similar to SCLC, a profile of *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}, and type II LCNECs with reduced expression of neuroendocrine genes and a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high} (Fig. 4).

Discussion

Here we provide the first comprehensive molecular analysis of LCNECs, which allowed distinguishing between two genomic subgroups with specific transcriptional patterns, defined as "type I LCNECs" and "type II LCNECs" (Fig. 4).

Type I and II LCNECs harbor key genomic alterations and oncogenic mutations, which are commonly found in SCLC, lung adenocarcinoma or squamous cell carcinoma (e.g., in *RAS* genes, *BRAF*, *NFE2L2*, as well as in *STK11* and *KEAP1* in the case of type I LCNECS, and *RB1* losses in the case of type II LCNECs). One possible explanation for this observation might be a high level of intra-tumor heterogeneity, combined with occurrence of two tumor types in a single tumor. However, the key alterations that we found in LCNECs were mostly clonal, with limited genomic intra-tumor heterogeneity. Furthermore, thorough comparisons of gene expression profiles did not suggest similarities between LCNECs and lung adenocarcinomas or squamous cell carcinomas. Thus, the combinations of distinct sets of mutations with specific patterns of gene expression supports the view that LCNECs are not a variant of the other types of lung cancer, but represent a distinct subgroup within the spectrum of neuroendocrine lung tumors.

In a more focused comparison with the most frequent neuroendocrine type of lung cancer, SCLC, type I LCNECs with *STK11* and *KEAP1* alterations exhibited a high degree of similarity with these carcinomas, as well as high expression of neuroendocrine genes and a profile of *ASCL1*^{high}/*DLL3*^{high}/*NOTCH*^{low}. By contrast, type II LCNECs with *RB1* alterations revealed reduced expression of neuroendocrine genes and a pattern of *ASCL1*^{low}/*DLL3*^{low}/*NOTCH*^{high}. Notch family members play a multifaceted role in the development of neuroendocrine tumors with cell-type specific tumor suppressor and oncogenic functions[40]. We have shown in earlier studies that *NOTCH* serves as a tumor suppressor in SCLC[6], which mostly harbor high-level expression of the negative regulator of Notch, *DLL3*[6,37,41] (Fig. 4). A recent clinical trial with an antibody-drug conjugate targeting the non-canonical inhibitory NOTCH ligand, Dll3, has shown early signs of clinical activity in SCLC[37,41]. We now demonstrate shared neuroendocrine pathways between SCLC and type I LCNECs, which may be similarly susceptible to this agent. On the other hand, type II LCNECs with alterations in *RB1* exhibited active Notch signaling (Fig. 4). Clinical trials have assessed the efficacy of an antibody targeting Notch 2 and 3 in SCLC, but recently failed in demonstrating a clinical benefit[42,43]. Therefore, future clinical trials involving therapeutics, targeting activating or inhibitory members of the Notch pathway will—in our view—require clear assignment of the respective molecular subtype.

Perhaps another noteworthy finding, type II LCNECs exhibited a pattern of gene expression with upregulation of immune related pathways (Fig. 3b, Fig. 4), which has similarly been observed in various other tumor types[28] and which may impact the response of patients to immunotherapy. Taken together, the precise distinction of high-grade neuroendocrine tumors representing as type I LCNECs and as *RB1*-mutated SCLC or type II LCNECs, may be pivotal to assess the efficacy of targeted therapeutics, including Notch pathway and immune checkpoint inhibitors.

Our sequencing studies did not reveal any somatic events that may cause the transcriptional discrepancy observed in LCNEC and SCLC tumors with *TP53* and *RB1* alteration, which raises the

question if all neuroendocrine tumors share the same cell of origin. It remains to be understood whether distinct tumor-specific cell of origins or cellular processes allow for plasticity and trans-differentiation that consequently lead to distinct molecular phenotypes. Importantly, histological trans-differentiation from lung adenocarcinoma to SCLC has been observed, both spontaneously or as resistance mechanisms to kinase inhibitors[44,45]; in some cases these were linked with a loss of *RB1*[4,46]. Previous studies involving genetically engineered mouse models and human cell lines have emphasized the phenomenon of transcriptional heterogeneity in SCLC and pointed to discordant expression of key lineage factors (e.g. *ASCL1*, *NEUROD1*, *REST*)[31,38]. By contrast, human primary tumors revealed a more complex expression pattern with co-expression of these transcriptional regulators. As a limitation of bulk tumor sequencing, advances in single cell sequencing may further aid to resolve and study the level of transcriptional intra-tumor heterogeneity in high-grade neuroendocrine tumors. While our studies pointed to transcriptional correlates of genomically defined subsets in LCNECs (type I and type II LNCECs), additional analyses on a larger dataset are warranted to further interrogate subcategories of high-grade neuroendocrine tumors.

In summary, we provide the first comprehensive characterization of neuroendocrine lung tumors, which integrates the molecular phenotypes of less frequent lung tumor subtypes. Despite the fact that LCNEC and SCLC tumors share some common clinical and histological characteristics, our study emphasizes pronounced differences in the pattern of genomic alterations and in their transcriptome profiles. The precise distinction of type I and type II LCNECs from SCLC is consequently pivotal to evaluate the response of patients to treatment options and to further understand morphological trans-differentiation processes in lung cancer patients.

Methods

Human specimens. The institutional review board (IRB) of the University of Cologne approved this study. Patient samples were obtained under IRB-approved protocols following written informed consent from all human participants. We collected and analyzed fresh-frozen samples of 75 LCNEC patients, which were provided by multiple collaborating institutions; 42 tumors were previously subject of other studies conducted by Rousseaux et al.[47] ($n = 25$) and Seidel et al.[7] ($n = 37$) (Supplementary Data 1). Clinical data were available for most patients, who were predominantly male (approximate ratio of 4:1) and current or former heavy smokers (Supplementary Data 1). All tumor samples were reviewed and confirmed by independent expert pathologists (E.B., W.T., and R.B.), and the diagnosis of LCNEC and the assessment of combined histological components were confirmed by H&E staining and immunohistochemistry, including markers for chromogranin A, synaptophysin, CD56 and Ki67. All tumors were positive for at least one neuroendocrine differentiation marker (Supplementary Data 1–2). Specimens containing >70% of tumor cells were processed for DNA and RNA extractions. DNA was extracted from matching normal material that was provided in the form of blood or adjacent non-tumorigenic lung tissue, which through pathological evaluation was confirmed to be free of tumor contaminants.

Nucleic acid extraction. Total DNA and RNA were obtained from fresh-frozen tumor tissue and matched fresh-frozen normal tissue or blood. Depending on the size of the tissue, 15–30 sections, each 20 µm thick, were cut using a cryostat (Leica) at −20 °C. The matched normal sample obtained from frozen tissue was processed the same way. Nineteen LCNEC cases were identified with mixed histological components of SCLC, lung adenocarcinomas and squamous cell carcinomas (Supplementary Data 1); in these cases nucleic acids were extracted from pure LCNEC regions by only dissecting the LCNEC component. DNA was extracted with the Gentra Puregene DNA extraction kit (Qiagen) and diluted to a working concentration of 100 ng/µL. The DNA was analyzed by agarose gel electrophoresis and confirmed to be of high-molecular weight (>10 kb). The DNA of tumor and normal material was confirmed to originate from the same patient by short tandem repeat (STR) analysis which was conducted at the Institute of Legal Medicine at the University of Cologne (Cologne, Germany), or by subsequent Affymetrix 6.0 SNP array and sequencing analyses.

RNA was isolated from tumor tissues by first lysing and homogenizing tissue sections with the Tissue Lyzer (Qiagen). The RNA was then extracted with the Qiagen RNAeasy Mini Kit. The RNA quality was analyzed at the Bioanalyzer 2100 DNA Chip 7500 (Agilent Technologies) and cases with a RNA integrity number (RIN) of over seven were considered for RNA-seq experiments.

Next-generation sequencing (NGS). WES was performed by first fragmenting 1 µg of DNA (Bioruptor, diagenode, Liége, Belgium). The DNA fragments were then end-repaired and adaptor-ligated with sample index barcodes. Following size selection, the SeqCap EZ Human Exome Library version 2.0 kit (Roche NimbleGen, Madison, WI, USA) was used to enrich for the whole exome. The DNA libraries were then sequenced with a paired-end 2 × 100 bp protocol aiming for an average coverage of 90× and 120× for the normal and tumor DNA, respectively. The primary data were filtered for signal purity with the Illumina Realtime Analysis software.

WGS was performed with a read length of 2 × 100 bp. The samples were processed to provide 110 Gb of sequence, thus amounting to a mean coverage of 30× for both tumor and matched normal.

For RNA-seq, cDNA libraries were prepared from PolyA + RNA following the Illumina TruSeq protocol for mRNA (Illumina, San Diego, CA, USA). The libraries were sequenced with a paired-end 2 × 100 bp protocol resulting in 8.5 Gb per sample, and thus in a 30× mean coverage of the annotated transcriptome.

Whole genome, whole exome and transcriptome sequencing reactions were performed on an Illumina HiSeq 2000 sequencing instrument (Illumina, San Diego, CA, USA).

Copy-number analysis by Affymetrix SNP 6.0 arrays. Human DNA from fresh-frozen tumors was analyzed with Affymetrix Genome-Wide Human SNP 6.0 arrays to determine copy-number alterations. Raw copy number data were computed by dividing tumor-derived signals by the mean signal intensities obtained from a subset of normal samples which were hybridized to the array in the same batch. Circular binary segmentation was applied to obtain segmented raw copy numbers[48]. Significant copy-number alterations were assessed with CGARS[49] at a threshold of $Q < 0.01$ (Supplementary Data 4).

Data processing and analyses of DNA sequencing data. The sequencing reads were aligned to the human reference genome NCBI build 37 (NCBI37/hg19) with BWA (version 0.6.1-r104)[50]. Possible PCR-duplicates were masked and not included for subsequent studies. We applied our in-house analysis pipeline[4,6,51] to analyze the data for somatic mutations, copy number alterations and genomic rearrangements. In brief, the mutation calling algorithm considers local sequencing depth, forward-reverse bias, and global sequencing error, to thus determine the presence of a mutated allele. We determined the somatic status of these mutations by assessing the absence of these variants in the sequencing data of the matched normal.

We determined genomic rearrangements from WGS data of 11 human LCNECs following the procedure as previously described[6,51]. In brief, the sequencing data were analyzed for discordant read-pairs, which were not within the expected mapping distance (>600 base pairs) or which revealed an incorrect orientation. Discordant read-pairs were analyzed for breakpoint-spanning reads, in which one read-pair shows partial alignments to two distinct genomic loci. Rearranged genomic loci were then reported at instances where at least one breakpoint-spanning read was identified. The genomic rearrangements called from each tumor sample were further filtered against the sequencing data of a matched normal and additionally against a library of normal genomes to thus minimize the detection of false-positive rearrangements.

Significantly mutated genes were analyzed as previously described[4,6]. In brief, we first determined the overall background mutation rate of each gene by computing its expected number of mutations assuming that all mutations are uniformly distributed across the genome. We also considered the ratio of synonymous to non-synonymous mutations into a combined statistical model to determine significantly mutated genes. Since mutation rates in non-expressed genes are often higher than the genome-wide background rate, we furthermore filtered for the expression of genes by referring to the transcriptome sequencing data of LCNECs. Only genes with a median FPKM (Fragments Per Kilobase Million) value of >1 in at least 35 out of 60 samples were considered (Methods section: RNA sequencing data processing and analyses). The significance of recurrently mutated genes was determined at a Q-value of <0.01 (Supplementary Data 7). Following previously described methods, we furthermore analyzed the data for significant enrichment of damaging mutations (including splice site, non-sense, and frameshift mutations)[6] and for significant clustering of mutations in genomic hotspots following a re-sampling based approach[4]. Significance was determined at a Q-value of 0.01, if the gene was affected in >10% of the samples (Supplementary Data 7). The damaging impact of mutations was further assessed by Polyphen[52].

The clonal status of mutations was assessed by computing for every mutation the "cancer cell fraction" (CCF), which defines within a tumor the fraction of cancer cells harboring that particular mutation[53]. The CCF was computed following our previously described approach[6]. In brief, this method first estimates tumor purity, ploidy, and absolute copy numbers, and computes for each mutation in a given sample the expected allele frequency under the assumption of clonality. The CCF is the quotient of the observed allelic fraction and the expected allelic fraction of a mutation. The distribution of CCFs for every mutation in a sample allowed to further identify distinct clusters and to thus assign the mutations to clonal and subclonal populations. The analysis described in Supplementary Fig. 2c considers mutations, which were assigned to clonal and subclonal fractions with a

probability >90%. In consideration of the sequencing coverage and the overall distribution of CCFs of every mutation in a sample, we furthermore determined the significant enrichment of mutations in a subclone at a *P*-value of 0.01 (Fig. 1b).

Mutational signatures analyses. Mutational signatures were analyzed in lung cancer subtypes applying previously described methods[54,55] and referring to the datasets of 77 lung adenocarcinomas (50 heavy-smokers (hs) and 27 non-smokers (ns) from the TCGA project)[2,25], 52 lung squamous cell carcinomas (from the TCGA project)[3], 109 SCLC[6], and 60 LCNECs from this study. Tumor cases with at least 30 somatic variants were selected and the list of variants were either extracted from Supplementary Materials[6] or COSMIC v68 (for the TCGA data)[20]. Variants were annotated with Annovar (version 12 Nov 2014). Gene strand orientations were retrieved from the RefSeqGene database using a customized Perl script. Variants were included in the analyses only if they could be successfully annotated. Single-base substitutions were classified into 96 types determined by the six possible substitutions (C:G > A:T, C:G > G:C, C:G > T:A, A:T > C:G, A:T > G:C, A:T > T:A) in their tri-nucleotides sequence context (16 combinations for each type of substitution). For extracting mutational signatures, we used the non-negative matrix factorization (NMF) algorithm developed by Lee et al.[56] and implemented in the Welcome Trust Sanger Institute (WTSI) mutational signatures framework.

Di-deoxynucleotide sequencing. Somatic alterations of interest were determined and confirmed by two independent sequencing approaches, which included WGS, WES, RNA-seq or di-deoxynucleotide sequencing. Di-deoxynucleotide chain termination sequencing (Sanger sequencing) was performed to validate mutations, genomic rearrangements, and chimeric fusion transcripts. Primer pairs were designed to amplify the target region encompassing the somatic alteration. The PCR reactions were performed either with genomic DNA or cDNA. The amplified products were subjected to Sanger sequencing and the respective electropherograms were analyzed by visual inspection using 4 Peaks or Geneious.

Analysis of RNA sequencing data. In order to detect chimeric transcripts, RNA-seq data were processed using TRUP[4,27]. In brief, paired-end RNA-seq reads were aligned to the human reference genome (NCBI37/hg19). We used TRUP to identify potential chimeric transcripts. Gene expression levels were determined with Cufflinks v2.0.2 referring only to paired-end reads that uniquely mapped within the expected mapping distance. The expression was quantified as FPKM (Fragments Per Kilobase Million) and the expression values served as a filter for identifying significantly mutated genes (Methods section: Data processing and analyses of DNA sequencing data).

Gene expression profiling and clustering studies. We analyzed transcriptome sequencing data from a total of $n = 341$ lung cancer samples. $N = 221$ samples referred to the data generated at the University of Cologne, Department of Translational Genomics, which included 41 lung adenocarcinoma[26,27], 61 pulmonary carcinoids[24], 53 SCLC[6], and 66 LCNECs from this present study. $N = 120$ samples were randomly selected from both the TCGA lung squamous cell carcinoma ($n = 60$)[3] and TCGA lung adenocarcinoma ($n = 60$) cohorts[2,25] referring to the Genomics Data Commons Legacy Archive. Sequencing data of lung adenocarcinomas from two different platforms aided in controlling for potential batch effects in subsequent studies. The raw sequencing reads of the RNA-seq data were all similarly processed to analyze for gene expression profiles. Sequencing reads which passed the quality control were mapped to the human reference genome (hg19) using MapSplice[57]. Picard Tools v1.64 (http://broadinstitute.github.io/picard/) was used to assess the alignment profile. SAMtools was used to sort and index the mapped reads and to determine transcriptome coordinates. The aligned reads were further filtered for indels, large inserts, and zero mapping quality with UBU v1.0 (https://github.com/mozack/ubu). RSEM[58], an expectation-maximization algorithm that refers to UCSC gene transcript and definitions, was applied to estimate transcript abundance. In order to allow comparisons between all RNA-Seq samples, raw RSEM read counts were normalized to the overall upper quartile[59]. The expression was quantified for 20,500 genes in 341 tumor samples and the median expression value was determined at RSEM = 209, which served as a reference threshold to classify for low and high expression. The expression determined by RSEM is provided for LCNECs in Supplementary Data 11.

For clustering purposes a set of genes that were both highly expressed and had highly variable expression patterns was identified in all lung cancer subtypes. Quality control procedures performed prior to any clustering analysis did not detect any evidence of batch effects.

After median centering the $\log_2(\text{RSEM} + 1)$ values by gene, unsupervised consensus clustering was applied using the ConsensusClusterPlus R package[60,61] with partitioning around medioids and a Spearman correlation-based distance. Additional hierarchical clustering of the consensus clustering classes was performed, applying average linkage and a Pearson correlation-based distance.

The statistical significance of the differences in gene expression patterns present in the subtype was assessed with the SigClust R package[62] by referring to the clustering gene sets and by using 1000 permutations and the default covariance estimation method. ClaNC[63] was used to identify genes whose expression patterns characterize the subtypes. R 3.0.2[61] was used to perform all statistical analyses and create all figures.

We first conducted consensus clustering of all lung cancer subtypes. The expression data of all lung cancer subtypes ($n = 341$) was analyzed and the 0.75 quantile of all $\log_2(\text{mean(RSEM)})$ values was used to identify highly expressed genes, while the 0.9 quantile of $\log_2(\text{variance(RSEM)})$ was used as a threshold to identify clustering gene sets that have highly variable expression patterns, which yielded a set of 1854 genes (Supplementary Fig. 6a). The samples were clustered with ConsensusClusterPlus following partition around medoids (PAM), and the ConsensusClusterPlus output along with gene expression heatmaps, principal components analysis, and silhouette plots was analyzed. Manual review of ConsensusClusterPlus output suggested a possible clustering solution based on $k = 6$ groups. However, two of the six groups included mainly lung adenocarcinoma samples and the gene expression heatmaps and PCA plots showed that these groups were quite similar. Thus, we chose to collapse these groups, thereby producing a five-class solution. The consensus clusters highly correlated with the histological subtypes as determined by Fisher's exact test Monte Carlo version ($P < 0.001$, 10,000 permutations): class A ($n = 66$; enriched for pulmonary carcinoids), class B ($n = 65$, enriched for lung squamous cell carcinomas), class C ($n = 108$, enriched for lung adenocarcinomas; data generated by different institutes), class D ($n = 38$, enriched for LCNEC and SCLC cases), and class E ($n = 64$, enriched for SCLC and LCNEC cases) (Supplementary Fig. 6b, Supplementary Data 12). ClaNC led to the identification of 875 classifier genes, which are displayed in the expression heatmaps (Fig. 2, Supplementary Fig. 6–7, Supplementary Data 13).

We then conducted consensus clustering of LCNECs, SCLC, lung adenocarcinomas, and squamous cell carcinomas. The unsupervised clustering approach was repeated for a subset of lung cancer subtypes; here excluding pulmonary carcinoids. The feature selection of highly variable (0.75 quantile) and highly expressed (0.9 quantile) genes across these lung tumor subtypes ($n = 280$) involved a gene set of 1855 genes and the consensus clustering process through hierarchical clustering suggested the presence of three expression clusters (expression subtypes): class A ($n = 98$, enriched for lung adenocarcinomas), class B ($n = 115$, enriched for LCNEC and SCLC), and class C ($n = 67$, enriched for lung squamous cell carcinomas). ClaNC identified 300 classifier genes which are displayed in the respective expression heatmaps (Supplementary Fig. 9).

We performed consensus clustering of LCNEC and SCLC through unsupervised clustering of the expression data of LCNEC and SCLC tumors alone ($n = 119$). Exploratory analyses of the gene expression data suggested the use of the 0.9 quantile of both the $\log_2(\text{mean(RSEM)})$ and $\log_2(\text{variance(RSEM)})$ values as thresholds for highly expressed and highly variably expressed genes. This produced a set of 1416 clustering genes. The Consensus clustering approach included hierarchical clustering and yielded four gene expression subtypes: class I ($n = 19$, only LCNECs), class II ($n = 49$, LCNEC and some SCLC tumors), class III ($n = 10$, SCLC and some LCNECs), and class IV ($n = 41$, mainly SCLC and some LCNECs) (Fig. 3, Supplementary Fig. 10–11, Supplementary Data 12). Hierarchical clustering of these cases revealed two main subgroups: one mainly formed by class I and II (enriched for LCNECs) and one mainly formed by class III and IV (enriched for SCLC) (Supplementary Fig. 11). 300 classifier genes were identified by ClaNC and are displayed in the expression heatmaps (Fig. 3, Supplementary Fig. 11, Supplementary Data 13).

We also performed consensus clustering of LCNECs with lung adenocarcinomas or lung squamous cell carcinomas. A gene set of (a) 1335 and (b) 1338 highly variable (0.85 quantile) and expressed genes (0.925 quantile) was identified in subsets of lung cancer tumors, including (a) LCNECs and lung adenocarcinomas ($n = 167$) and (b) LCNECs and lung squamous cell carcinomas ($n = 126$). The consensus clustering approach through PAM (partitioning around medoids) suggested in both cases two transcriptional subclasses: for approach (a) class A ($n = 70$, mainly LCNECs) and class B ($n = 97$, mainly lung adenocarcinomas); and for approach (b) class A ($n = 58$, mainly LCNECs) and class B ($n = 68$, mainly lung squamous cell carcinomas). ClaNC identified 100 classifier genes in each approach, which were used for the expression heatmaps (Supplementary Fig. 9).

We furthermore performed consensus clustering of LCNECs alone. The transcriptional data on LCNECs was analyzed and hierarchical clustering referred to 475 very highly expressed (0.875 quantile) and very highly variable (0.975 quantile) genes. The consensus clustering approach yielded a $k = 4$ clustering solution: class 1 ($n = 11$), class 2 ($n = 21$), class 3 ($n = 24$), and class 4 ($n = 10$). ClaNC was then applied to the clustering solution, which further identified 540 classifier genes (Supplementary Fig. 13, Supplementary Data 13).

Differential expression analysis. The SAMR R package[64] was used to identify genes that were differentially expressed in the expression subtypes using 1000 permutations and a *Q*-value threshold of 0.05 (Supplementary Data 13). We then used the DAVID annotation database[65,66] to identify pathways that were enriched for differentially expressed genes at $P < 0.0001$ (Supplementary Data 14).

Immunohistochemistry. FFPE tissue sections of 3-µm thickness were stained for hematoxylin and eosin (H&E) and immunohistochemistry (IHC) was conducted for CD56 (*NCAM1*), Synaptophysin (*SYP*), Chromogranin A (*CHGA*, clone DAK-A3), TTF-1 (*NKX2-1*, clone 8G7G3/1), and Rb1 (*RB1*, clone 1F8 (ab81701; Abcam,

Cambridge, UK) (Supplementary Data 2, Supplementary Table 1). Hematoxylin and eosin (H&E) were scanned and can be viewed online or with the Pannoramic Viewer software (3D Histech) as specified in Supplementary Data 2 (for further information see "Data Availability").

Specifically, IHC for Rb1 was performed with the Novolink max polymer detection system (RE7280-CE, Leica Biosystems, Wetzlar, Germany) using EDTA buffer pH 8.0 (K038, Diagnostic BioSystems, Pleasanton, USA) antigen retrieval (4 × 5 min by microwave 700 W). The primary antibody was incubated overnight at 4 °C; the secondary antibody was incubated for 30 min at room temperature. The signal was visualized by diaminobenzidine after incubation for 5 min at room temperature. Sections were counter-stained with hematoxylin for 5 min. The *H*-score method was used for evaluating the immunostaining with Rb1 by multiplying the intensity of the staining (0: no staining, 1: weak, 2: moderate and 3: strong staining) with the percentage of the tumor or stroma stained. The minimum score was 0 and the maximum was 300 (Supplementary Data 2).

Fluorescence in situ hybridization assay. Genomic rearrangements of *PTK6* on chromosome 20 were assessed through a dual-color break-apart fluorescence in situ hybridization (FISH) assay following previous protocols[67]. In brief, the BAC clone RP11-939M14 labeled centromeres with biotin (red signal) and CTD-3228E10 labeled telomeric sites with digoxigenin (green signal). The samples were analyzed with a 63× oil immersion objective at a fluorescence microscope (Zeiss, Jena, Germany) equipped with appropriate filters, a charge-coupled device camera and the FISH imaging and capturing software Metafer 4 (Metasystems, Altlussheim, Germany). Two independent scientists analyzed the experiment (R.M. and S.P.). Translocations were derived from a split of a signal pair, resulting in a single red and green signal, single red or green signals resulting from signal loss, were referred to as a rearrangement through deletion. In cases where cells were wild type and displayed no rearrangements, a juxtaposed red and green signal (mostly forming a yellow signal) was observed.

NTRK1 break-apart FISH were performed with the ZytoLight SPEC *NTRK1* Dual Color Break Apart Probe (ZytoVision, Bremerhaven, Germany). According to previous protocols[68], 4 µm sections of FFPE tissue were treated with the Paraffin pretreatment reagent kit (Vysis, Abbott Molecular), and then stained with the probes following the instructions of the manufacturer. An *NTRK1* rearrangement was diagnosed when >15% of the nuclei showed either a split pattern with 3′ and 5′ signals separated by a distance superior to the diameter of the largest signal, or isolated 3′ (orange) signals.

Data availability. Sequencing data and Affymetrix 6.0 SNP array data are deposited at the European Genome-phenome Archive, which is hosted by the EBI (EGA, http://www.ebi.ac.uk/ega/), under accession number EGAS00001000708. Histological images of FFPE samples from LCNECs of this study are deposited as H&E images (domain 1: https://teleslide.chu-grenoble.fr/ > acces libre > recherche > recherche/TP/LCNEC-study > code access 1793) or as data files compatible with the Pannoramic Viewer software (3D Histech) (domain 2: https://uni-koeln.sciebo.de/index.php/s/xMjs4dqJpqbOVDn); an overview is provided in Supplementary Data 2.

Received: 10 April 2017 Accepted: 18 January 2018
Published online: 13 March 2018

References

1. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
2. Collisson, Ea et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
3. Hammerman, P. S. et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
4. Peifer, M. et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **44**, 1104–1110 (2012).
5. Rudin, C. M. et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.* **44**, 1111–1116 (2012).
6. George, J. et al. Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).
7. Seidel, D. A genomics-based classification of human lung tumors. *Sci. Transl. Med.* **5**, 209ra153 (2013).
8. Bhattacharjee, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA* **98**, 13790–13795 (2001).
9. Hayes, D. N. et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24**, 5079–5090 (2006).
10. Wilkerson, M. D. et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* **16**, 4864–4875 (2010).
11. Chen, F. et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene* **36**, 1384–1393 (2017).
12. Travis, W. D. Advances in neuroendocrine lung tumors. *Ann. Oncol.* **21**, vii65–71 (2010).
13. Travis, W. D. et al. The 2015 World Health Organization Classification of lung tumors. *J. Thorac. Oncol.* **10**, 1243–1260 (2015).
14. Fasano, M. et al. Pulmonary large-cell neuroendocrine carcinoma: from epidemiology to therapy. *J. Thorac. Oncol.* **10**, 1133–1141 (2015).
15. Karlsson, A., Brunnström, H., Lindquist, K. E. & Jirström, K. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer Patient material. *Oncotarget* **6**, 22028–22037 (2015).
16. Rekhtman, N. et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma-like and non-small cell carcinoma-like subsets. *Clin. Cancer Res.* **22**, 3618–3629 (2016).
17. Miyoshi, T. et al. Genomic profiling of large-cell neuroendocrine carcinoma of the lung. *Clin. Cancer Res.* **23**, 757–765 (2017).
18. Jones, M. H. et al. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* **363**, 775–781 (2004).
19. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
20. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2014).
21. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
22. Weiss, J. et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci. Transl. Med.* **2**, 62ra93 (2010).
23. Wistuba, I. I., Gazdar, A. F. & Minna, J. D. Molecular genetics of small cell lung carcinoma. *Semin. Oncol.* **28**, 3–13 (2001).
24. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodeling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
25. Imielinski, M. et al. Mapping the Hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
26. Fernandez-Cuesta, L. et al. CD74-NRG1 fusions in lung adenocarcinoma. *Cancer Discov.* **4**, 415–422 (2014).
27. Fernandez-Cuesta, L. et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* **16**, 7 (2015).
28. Rooney, M. S., Shukla, Sa, Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
29. Augustyn, A. et al. ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung cancers. *Proc. Natl Acad. Sci. USA* **111**, 14788–14793 (2014).
30. Westerman, B. A. et al. Basic helix-loop-helix transcription factor profiling of lung tumors shows aberrant expression of the proneural gene atonal homolog 1 (ATOH1, HATH1, MATH1) in neuroendocrine tumors. *Int. J. Biol. Markers* **22**, 114–123 (2007).
31. Borromeo, M. D. et al. ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. *Cell Rep.* **16**, 1259–1272 (2016).
32. Sutherland, K. D. et al. Cell of origin of small cell lung cancer: inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. *Cancer Cell* **19**, 754–764 (2011).
33. Park, K. et al. Characterization of the cell of origin for small cell lung cancer. *Cell Cycle* **10**, 2806–2815 (2011).
34. Song, H. et al. Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. *Proc. Natl Acad. Sci. USA* **109**, 17531–17536 (2012).
35. Sugano, M., Nagasaka, T. & Sasaki, E. HNF4 a as a marker for invasive mucinous adenocarcinoma of the lung. *Am. J. Surg. Pathol.* **37**, 211–218 (2013).
36. Snyder, E. L. et al. Article Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma. *Mol. Cell* **50**, 185–199 (2013).
37. Saunders, L. R. et al. A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. *Sci. Transl. Med* **7**, 302ra136 (2015).
38. Lim, J. S. et al. Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature* **545**, 360–364 (2017).

39. Kazarian, M. & Laird-Offringa, Ia Small-cell lung cancer-associated autoantibodies: potential applications to cancer diagnosis, early detection, and therapy. *Mol. Cancer* **10**, 33 (2011).
40. Ranganathan, P., Weaver, K. L. & Capobianco, A. J. Notch signalling in solid tumours: a little bit of everything but not all the time. *Nat. Rev. Cancer* **11**, 338–351 (2011).
41. Pietanza, M. C. et al. Safety, activity, and response durability assessment of single agent rovalpituzumab tesirine, a delta-like protein 3 (DLL3)-targeted antibody drug conjugate (ADC), in small cell lung cancer (SCLC). *Eur. J. Cancer.* **51**, S712 (2015).
42. Yen, W. C. et al. Targeting notch signaling with a Notch2/Notch3 antagonist (Tarextumab) inhibits tumor growth and decreases tumor-initiating cell frequency. *Clin. Cancer Res.* **21**, 2084–2095 (2015).
43. Pietanza, M. C. et al. Final results of phase Ib of tarextumab (TRXT, OMP-59R5, anti-Notch2/3) in combination with etoposide and platinum (EP) in patients (pts) with untreated extensive-stage small-cell lung cancer (ED-SCLC). *J. Clin. Oncol.* 33, 7508 (2015).
44. Zakowski, M. F., Ladanyi, M. & Kris, M. G. EGFR mutations in small-cell lung cancers. *N. Engl. J. Med.* **355**, 213–215 (2006).
45. Morinaga, R. et al. Sequential occurrence of non-small cell and small cell lung cancer with the same EGFR mutation. *Lung Cancer* **58**, 411–413 (2007).
46. Niederst, M. J. et al. RB loss in resistant EGFR mutant lung adenocarcinomas that transform to small-cell lung cancer. *Nat. Commun.* **6**, 6377 (2015).
47. Rousseaux, S. et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung. *Cancers* **5**, 1–12 (2013).
48. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
49. Lu, X., Thomas, R. K. & Peifer, M. CGARS: cancer genome analysis by rank sums. *Bioinformatics* **30**, 1295–1296 (2014).
50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
51. Fernandez-Cuesta, L. et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).
52. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations a. *Nature* **7**, 248–249 (2010).
53. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
54. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
55. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
56. Lee, S. Y., Song, H. A. & Amari, S. I. A new discriminant NMF algorithm and its application to the extraction of subtle emotional differences in speech. *Cogn. Neurodyn.* **6**, 525–535 (2012).
57. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, 1–14 (2010).
58. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
59. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinforma.* **11**, 94 (2010).
60. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
61. R Core Team, R. F. for S. C. *R: A language and environment for statistical computing*. (2014). Available at http://www.r-project.org/
62. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical significance of clustering for high-dimension, low–sample size dataset. *J. Am. Stat. Assoc.* **103**, 1281–1293 (2008).
63. Dabney, A. R. Classification of microarrays to nearest centroids. *Bioinformatics* **21**, 4148–4154 (2005).
64. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
65. Huang, D. W. & Lempicki, R. A. & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
66. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
67. Menon, R. et al. Somatic copy number alterations by whole-exome sequencing implicates YWHAZ and PTK2 in castration-resistant prostate cancer. *J. Pathol.* **231**, 505–516 (2013).
68. McLeer-Florin, A. et al. Dual IHC and FISH testing for ALK gene rearrangement in lung adenocarcinomas in a routine practice. *J. Thorac. Oncol.* **7**, 348–354 (2012).

Acknowledgements

This work was supported by the German Cancer Aid (Deutsche Krebshilfe) as part of the small cell lung cancer genome sequencing consortium (grant ID: 109679 to R.K.T., M.P., R.B., P.N., M.V., and S.A.H.), by the German Ministry of Science and Education (BMBF) as part of the e:Med program (grant no. 01ZX1303A to R.K.T., R.B., U.L., M.P. and J. W, and grant no. 01ZX1406 to M.P.), by the EU-Framework program CURELUNG (HEALTH-F2-2010-258677 to R.K.T., J.W., E.B., and L.R.), by the Deutsche Forschungsgemeinschaft (DFG; through TH1386/3-1 to R.K.T.), by the Deutsche Krebshilfe as part of the Oncology Centers of Excellence funding program (R.K.T.), by the National Institute of Health (NIH U10CA181009 to D.N.H.), by the German Cancer Consortium (DKTK) Joint Funding program, by Associazione Italiana per la Ricerca sul Cancro (AIRC, IG 16847 to L.R.), and by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396 (L.B.A.). J. G. received funding as part of the IASLC Young Investigator award. L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. We are indebted to the patients donating their tumor specimens as part of the Clinical Lung Cancer Genome Project initiative. We thank the regional computing center of the University of Cologne (RRZK) for providing the CPU time on the DFG-funded supercomputer 'CHEOPS', as well as the support. We would like to acknowledge that Australian specimens were provided with assistance of the Victorian Cancer Biobank. We furthermore thank Johannes Berg, Chau Nguyen, Philipp Lorimier, Elisabeth Kirst, and César Tejerina Álvarez for their technical assistance.

Author contributions

R.K.T., L.F.-C., J.G., and E.B. conceived and designed the project. J.G., V.W., S.P., S.A.H., M.F., D.N.H., W.D.T., L.F.C., E.B., and R.K.T. supervised the work and gave scientific input. J.G., V.W., L.B.A., L.M., T.M.D., M.A., No.L., M.P., G.B., R.S., A.D.R., M.G.S., M. D.W., S.A.H., M.O., Y.C., and L.F.-C. performed computational and statistical analyses. A.M.F., F.M., R.M., F.L., C.M., I.D., D.S., P.S., J.A., C.B., and M.B. performed experiments. R.B., W.D.T., and E.B. performed pathological review. D.M.S., C.G.B., S.L., A.S., W.W., V.T., O.T.B., M.L.I., A.H., S.S., S.A., G.W., B.S., L.R., U.P., I.P., J.H.C., J.S., R.B., Ni. L., and E.B. contributed with samples. D.S., G.B., F.L., L.M., Ni.L., V.A., U.L., P.N., P.M. S., J.D.M., J.W., M.V., and T.Z. helped with logistics. J.G., R.K.T., and L.F.-C. wrote the manuscript, which was reviewed by all the co-authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03099-x.

Competing interests: L.F.-C. and R.K.T. are inventors on a patent application related to findings described in this manuscript. R.K.T. is a founder of NEO New Oncology GmbH, now part of Siemens Healthcare. R.K.T. received consulting and lecture fees (Merck, Roche, Lilly, Boehringer Ingelheim, AstraZeneca, Daiichi-Sankyo, MSD, NEO New Oncology, Puma, Clovis). R.B. is a cofounder and owner of Targos Molecular Diagnostics and received honoraria for consulting and lecturing from AstraZeneca, Boehringer Ingelheim, Merck, Roche, Novartis, Lilly, and Pfizer. J.W. received consulting and lecture fees from Roche, Novartis, Boehringer Ingelheim, AstraZeneca, Bayer, Lilly, Merck, Amgen and research support from Roche, Bayer, Novartis, Boehringer Ingelheim. T.Z. received honoraria from Roche, Novartis, Boehringer Ingelheim, Lilly, Merck, Amgen and research support from Novartis. B.S. received consulting fees from AstraZeneca, Roche-Genentech, Pfizer, Novartis, Merck, and Bristol Myers Squibb. The remaining authors declare no competing financial interest.

Julie George[1], Vonn Walter[2,3], Martin Peifer[1,4], Ludmil B. Alexandrov[5], Danila Seidel[1], Frauke Leenders[1], Lukas Maas[1], Christian Müller[1], Ilona Dahmen[1], Tiffany M. Delhomme[6], Maude Ardin[7], Noemie Leblay[6], Graham Byrnes[8], Ruping Sun[9], Aurélien De Reynies[10], Anne McLeer-Florin[11], Graziella Bosco[1], Florian Malchers[1], Roopika Menon[12], Janine Altmüller[4,13,14], Christian Becker[13], Peter Nürnberg[4,13,15], Viktor Achter[16], Ulrich Lang[16,17], Peter M. Schneider[18], Magdalena Bogus[18], Matthew G. Soloway[2], Matthew D. Wilkerson[19], Yupeng Cun[1,4], James D. McKay[6], Denis Moro-Sibilot[20], Christian G. Brambilla[20], Sylvie Lantuejoul[21,22], Nicolas Lemaitre[21], Alex Soltermann[23], Walter Weder[24], Verena Tischler[23], Odd Terje Brustugun[25,26], Marius Lund-Iversen[27], Åslaug Helland[24,25], Steinar Solberg[28], Sascha Ansén[29], Gavin Wright[30], Benjamin Solomon[31], Luca Roz[32], Ugo Pastorino[33], Iver Petersen[34], Joachim H. Clement[35], Jörg Sänger[36], Jürgen Wolf[29], Martin Vingron[9], Thomas Zander[37,38], Sven Perner[39], William D. Travis[40], Stefan A. Haas[9], Magali Olivier[7], Matthieu Foll[6], Reinhard Büttner[38], David Neil Hayes[2], Elisabeth Brambilla[21], Lynnette Fernandez-Cuesta[1,6] & Roman K. Thomas[1,38,41]

[1]Department of Translational Genomics, Center of Integrated Oncology Cologne-Bonn, Medical Faculty, University of Cologne, Cologne, 50931, Germany. [2]UNC Lineberger Comprehensive Cancer Center School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7295, USA. [3]Department of Biochemistry and Molecular Biology, Penn State Milton S. Hershey Medical Center, 500 University Drive, Hershey, PA 17033, USA. [4]Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50931 Cologne, Germany. [5]Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, La Jolla, San Diego, CA 92093, USA. [6]Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer (IARC-WHO), Lyon, 69008, France. [7]Molecular Mechanisms and Biomarkers Group, Section of Mechanisms of Carcinogenesis, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France. [8]Section of Environment and Radiation, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France. [9]Computational Molecular Biology Group, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany. [10]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, 14 rue Corvisart, Paris, 75013, France. [11]CHU Grenoble Alpes, UGA/INSERM U1209/CNRS, Grenoble, France. [12]NEO New Oncology GmbH, 51105 Cologne, Germany. [13]Cologne Center for Genomics (CCG), University of Cologne, 50931 Cologne, Germany. [14]Institute of Human Genetics, University Hospital Cologne, 50931 Cologne, Germany. [15]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, 50931 Cologne, Germany. [16]Computing Center, University of Cologne, 50931 Cologne, Germany. [17]Department of Informatics, University of Cologne, 50931 Cologne, Germany. [18]Institute of Legal Medicine, University Hospital Cologne, 50823 Cologne, Germany. [19]Department of Genetics, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, NC, 27599-7295, USA. [20]CHUGA Grenoble, INSERM U 1209, University Grenoble Alpes, Institute of Advanced Biosciences (IAB), 38043, CS10217 Grenoble, France. [21]Department of Pathology, CHUGA, INSERM U 1209, University of Grenobles Alpes, Institute of Advanced Biosciences (IAB), 38043, CS10217 Grenoble, France. [22]Department of Biopathology, Centre Léon Bérard UNICANCER, 69008 Lyon, France. [23]Institute of Pathology and Molecular Pathology, University Hospital Zurich, 8091 Zurich, Switzerland. [24]Department of Thoracic Surgery, University Hospital Zurich, 8091 Zurich, Switzerland. [25]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, N-0424 Oslo, Norway. [26]Department of Oncology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway. [27]Department of Pathology, Norwegian Radium Hospital, Oslo University Hospital, N-0310 Oslo, Norway. [28]Department of Thoracic Surgery, Rikshospitalet, Oslo University Hospital, N-0027 Oslo, Norway. [29]Department of Internal Medicine, Center of Integrated Oncology Cologne-Bonn, University Hospital Cologne, 50937 Cologne, Germany. [30]Department of Surgery, St. Vincent's Hospital, Peter MacCallum Cancer Centre, 3065 Melbourne, Victoria, Australia. [31]Department of Haematology and Medical Oncology, Peter MacCallum Cancer Centre, 3065 Melbourne, Victoria, Australia. [32]Tumor Genomics Unit, Department of Experimental Oncology and Molecular Medicine, Fondazione IRCCS—Istituto Nazionale Tumori, Via Venezian 1, 20133 Milan, Italy. [33]Thoracic Surgery Unit, Department of Surgery, Fondazione IRCCS Istituto Nazionale Tumori, 20133 Milan, Italy. [34]Institute of Pathology, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany. [35]Department of Internal Medicine II, Jena University Hospital, Friedrich-Schiller-University, 07743 Jena, Germany. [36]Institute for Pathology Bad Berka, 99438 Bad Berka, Germany. [37]Gastrointestinal Cancer Group Cologne, Center of Integrated Oncology Cologne-Bonn, Department I for Internal Medicine, University Hospital of Cologne, 50823 Cologne, Germany. [38]Department of Pathology, University Hospital Cologne, 50937 Cologne, Germany. [39]Pathology of the University Medical Center Schleswig-Holstein, Campus Luebeck and the Research Center Borstel, Leibniz Center for Medicine and Biosciences, 23538 Luebeck and 23845 Borstel, Borstel, Germany. [40]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA. [41]German Cancer Research Center, German Cancer Consortium (DKTK), 69120 Heidelberg, Germany. These authors contributed equally: Julie George, Vonn Walter, Lynnette Fernandez-Cuesta.

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

Genome and Epigenome

Cancer Research

Integrative Genomic Characterization Identifies Molecular Subtypes of Lung Carcinoids

Saurabh V. Laddha[1], Edaise M. da Silva[2], Kenneth Robzyk[2], Brian R. Untch[3], Hua Ke[1], Natasha Rekhtman[2], John T. Poirier[4], William D. Travis[2], Laura H. Tang[2], and Chang S. Chan[1,5]

Abstract

Lung carcinoids (LC) are rare and slow growing primary lung neuroendocrine tumors. We performed targeted exome sequencing, mRNA sequencing, and DNA methylation array analysis on macro-dissected LCs. Recurrent mutations were enriched for genes involved in covalent histone modification/chromatin remodeling (34.5%; *MEN1*, *ARID1A*, *KMT2C*, and *KMT2A*) as well as DNA repair (17.2%) pathways. Unsupervised clustering and principle component analysis on gene expression and DNA methylation profiles showed three robust molecular subtypes (LC1, LC2, LC3) with distinct clinical features. *MEN1* gene mutations were found to be exclusively enriched in the LC2 subtype. LC1 and LC3 subtypes were predominately found at peripheral and endobronchial lung, respectively. The LC3 subtype was diagnosed at a younger age than LC1 and LC2 subtypes. IHC staining of two biomarkers, ASCL1 and S100, sufficiently stratified the three subtypes. This molecular classification of LCs into three subtypes may facilitate understanding of their molecular mechanisms and improve diagnosis and clinical management.

Significance: Integrative genomic analysis of lung carcinoids identifies three novel molecular subtypes with distinct clinical features and provides insight into their distinctive molecular signatures of tumorigenesis, diagnosis, and prognosis.

Introduction

Lung carcinoids (LC) are an indolent and rare type of primary lung neoplasms that are, in general, understudied. The 2015 World Health Organization (WHO; ref. 1) classification of LCs includes atypical carcinoids (AC; ~0.2% prevalence) and typical carcinoids (TC; ~2% prevalence). TCs are slow growing tumors that rarely spread beyond the lungs whereas ACs are faster growing tumors and have a greater chance of metastasizing to other tissues (2). The WHO classification relies mainly on morphology, proliferation rate (mitotic index), and necrosis assessment (3). This current method of classification has its drawbacks as studies have shown that the reproducibility of cancer classification and its prognostic efficacy have high interobserver variability (3, 4), especially for differentiating between TC and AC (5). Recent WHO classifications highlight use of the Ki67 cell proliferation marker to distinguish ACs from TCs (1). However, overlapping distribution of Ki67 between ACs and TCs does not enable reliable stratification between well-differentiated LCs (6, 7). It has also been reported that TCs and ACs are overdiagnosed as small cell lung carcinomas (SCLC) in small crush biopsy specimens (8), a situation where artifacts in specimens appear as bluish clusters in which cellular details are not recognizable. As SCLCs are highly malignant, incorrect diagnosis of TC and AC tumors as SCLC can subject patients to unnecessary stress and treatment (8). More accurate molecular diagnostic tools and stratification for LCs will help ensure more appropriate treatment and clinical management.

Previous genomic analysis of LC tumors has identified recurrent mutations in *MEN1*, *PSIP1*, and *ARID1A* (9), whereas no significant mutations or focal copy alterations were observed in genes that are frequently mutated in non–small cell lung cancer (NSCLC), large cell neuroendocrine carcinoma (LCNEC), and SCLC (e.g., *KRAS*, *TP53*, *EGFR*, and *RB1*; ref. 10). The different mutation spectrum and low mutation burden (9) of LCs indicate they are distinct from NSCLC and high-grade lung neuroendocrine tumors (NET). It is not known if there are distinct molecular subtypes of LCs or their cells of origin.

In this study, we performed genotyping on 29 LCs to detect mutations in a 354-cancer gene panel, mRNA sequencing ($n = 30$) and DNA methylation 450K-array analysis ($n = 18$) and identified 3 molecular subtypes with distinct clinical features. We also identified 2 key biomarkers (*ASCL1* and *S100*) to stratify these subtypes. Integration of genetic and epigenetic signatures distinguishes each subtype of carcinoid, providing deeper insight into their distinctive molecular signatures of tumorigenesis as well as cells of origin.

Materials and Methods

Patient's data

Retrospective and prospective reviews of 30 LC neoplasms were performed using the pathology files and institutional database at

[1]Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, New Jersey. [2]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York. [3]Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York. [4]Thoracic Oncology Service, Memorial Sloan-Kettering Cancer Center, New York, New York. [5]Department of Medicine, Rutgers Robert Wood Johnson Medical School, New Brunswick, New Jersey.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Authors: Chang S. Chan, Rutgers Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08903. Phone: 732-235-7363; E-mail: chanc3@cinj.rutgers.edu; and Laura H. Tang, Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065. Phone: 212-639-5905. E-mail: tangl@mskcc.org

Cancer Res 2019;79:4339–47

doi: 10.1158/0008-5472.CAN-19-0214

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

Memorial Sloan Kettering Cancer Center (MSKCC, New York, NY). All studies were conducted in accordance with appropriate ethical guidelines (following U.S. Common Rule) and with Institutional Review Board approval. Written informed consent was obtained from the patients. All patients were evaluated clinically with confirmed pathologic diagnoses, appropriate radiological and laboratory studies, and surgical or oncological management. Fresh-frozen tumor and paired normal tissues were obtained from MSKCC's tissue bank under Institutional Review Board protocol. Targeted cancer gene panel DNA sequencing, RNA sequencing (RNA-seq), and DNA methylation array were performed on fresh-frozen samples. Relevant clinical and pathologic information is presented in Supplementary File S1.

DNA sequencing and analysis

DNA extraction from microdissected tumor samples and normal adjacent tissues was performed using a commercially available DNA Extraction Kit (DNeasy Blood and Tissue Kit, Catalog No. 69504; Qiagen), according to the manufacturer's protocol. Targeted sequencing on 29 LCs was performed using MSK-IMPACT (11) hybrid capture cancer gene panel ($n = 354$). Single-nucleotide variants and short indels (<30 bp) were identified and annotated using MSK-IMPACT pipeline (11). Briefly, raw reads were filtered based on quality, mapped to NCBI b37 genome using BWA–MEM version 0.7.5a (http://arxiv.org/abs/1303.3997), post-processed using GATK (12), and variant identification using MuTect (13). Variants were filtered based on its entry in NCBI-dbSNPs (http://www.ncbi.nlm.nih.gov/snp), 1000G project (http://www.1000genomes.org/), and COSMIC (http://cancer.sanger.ac.uk/cosmic). Filtered variants were manually reviewed on IGV. We created the mutational OncoPrint plot for our 29 LCs dataset using the online cBioPortal website (http://www.cbioportal.org/oncoprinter.jsp).

RNA-seq and data analysis

Total RNA extraction from microdissected frozen tumor samples was performed using RNeasy Mini Kit (Catalog No. 74106) followed by Illumina HiSeq sequencing (2×100 bp). We performed RNA-seq data analysis as described previously (14, 15). Briefly, raw fastq files were examined for sequencing quality control using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Sequencing reads were mapped to human transcripts corresponding to hg19 Genepattern (16) GTF annotations using RSEM (17) with default parameters. STAR (18) aligner was used to map reads in RSEM algorithm followed by calculating gene expression from mapped BAM files. Gene transcripts mapped data were normalized to transcript per million (TPM). We used log2 (TPM+1) values for all downstream analysis. For unsupervised clustering, we used Pearson distance metric and hclust (ward.D2) method (unless stated otherwise). PCA was done using *prcomp* in R. R (http://www.r-project.org/) was used for all analysis and visualization of data. The R package DeSeq2 (19) was used to identify differentially expressed genes using Benjamini and Hochberg corrected *P* value (<0.05) and fold change of ≥ 2. We used DAVID bioinformatics resources (20) version 6.8 and gene set enrichment analysis (GSEA; including PreRanked test; ref. 21) to find significant pathways, gene ontology terms and transcription factor motif analysis with default parameter.

Independent dataset of LCs

To validate our novel molecular subtyping, we used gene expression and mutation data from an independent LC ($n = 65$) dataset. The gene expression and mutational data were downloaded from supplement data files (https://www.nature.com/articles/ncomms4518#supplementary-information) reported in ref. 9. This gene expression data were reported as transcript expression instead of gene expression. We used callapseRows (22) on transcript expression to convert to respective gene expression using MaxVariance option. Gene expression for LCs signature (top 100 variant genes from our 30 LCs dataset) was extracted from this RNAseq dataset. Unsupervised clustering and PCA analysis on this dataset were performed as previously described for our 30 LCs dataset.

DNA methylation and analysis

DNA were extracted from LC samples and interrogated for CpG DNA methylation using the Illumina 450K array platform (Illumina Inc.). For CpG DNA methylation data analysis, we used ChAMP (23) package in R with default parameters. Briefly, IDAT raw data files were imported in R and filtered to exclude samples with detection *P*-value <0.01 and bead count <3 in at least 5% of samples and normalized using FunctionNormalization (24). Additional filtering was performed to remove probes that have annotated SNPs, present on X and Y chromosome, and have CpG probes mapped to multiple locations. β values for 413,176 probes were used for all subsequent analysis. Subtype specific differentially methylated CpG probes (DMP) and CpG island were identified using COHCAP (default parameter; ref. 25). For DMP, we focused on probes present at TSS1500/200 and first exon for subsequent analysis. We integrated gene expression and DNA methylation to investigate subtype-specific gene expression using default parameter for COHCAP.integrate.avg.by.island with FDR *P* value <0.01.

IHC staining

IHC staining was performed using commercially available antibodies at optimal dilutions as follows: ASCL1 (a-MASH1; monoclonal, 1:300; BD Biosciences) and S100 (monoclonal, 1:4,000; BG). Briefly, IHC was performed by a standard protocol on Ventana Discovery XT automated stainer (Ventana Medical Systems Inc.) for the S100 antibody. Antigen retrieval was performed with Cell Conditioning 1 buffer (CC1; citrate buffer pH 6.0, Ventana). For the ASCL1 (a-MASH1) antibody, ER2 pretreatment was used and the Leica Bondmax autostainer (Leica). Immunoreactivity was performed on whole tissue sections of formalin-fixed and paraffin-embedded tissue section as well as on tissue microarray (TMA) constructed from 173 pulmonary carcinoid tumors formalin-fixed, paraffin-embedded tumor specimens were used for TMA construction. Briefly, 4 representative tumor areas were marked on H&E-stained slides, and cylindrical 0.6-mm tissue cores were arrayed from the corresponding paraffin blocks into a recipient block using an automated tissue arrayer ATA-27 (Beecher Instruments). From each TMA, 4-μm-thick paraffin sections were prepared for IHC. In all, 173 cases with adequate cores were available for IHC analysis (26). The results were semiquantitatively scored based on the percentage of reactive tumor cells ($\geq$25% tumor cells) and the intensity of staining.

Data availability

The authors declare that all data supporting the findings of this study are available within the article and its Supplementary data and figures. Data generated in this study were deposited to NCBI under GEO SuperSeries GSE118133 (GSE118131 for RNAseq and

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

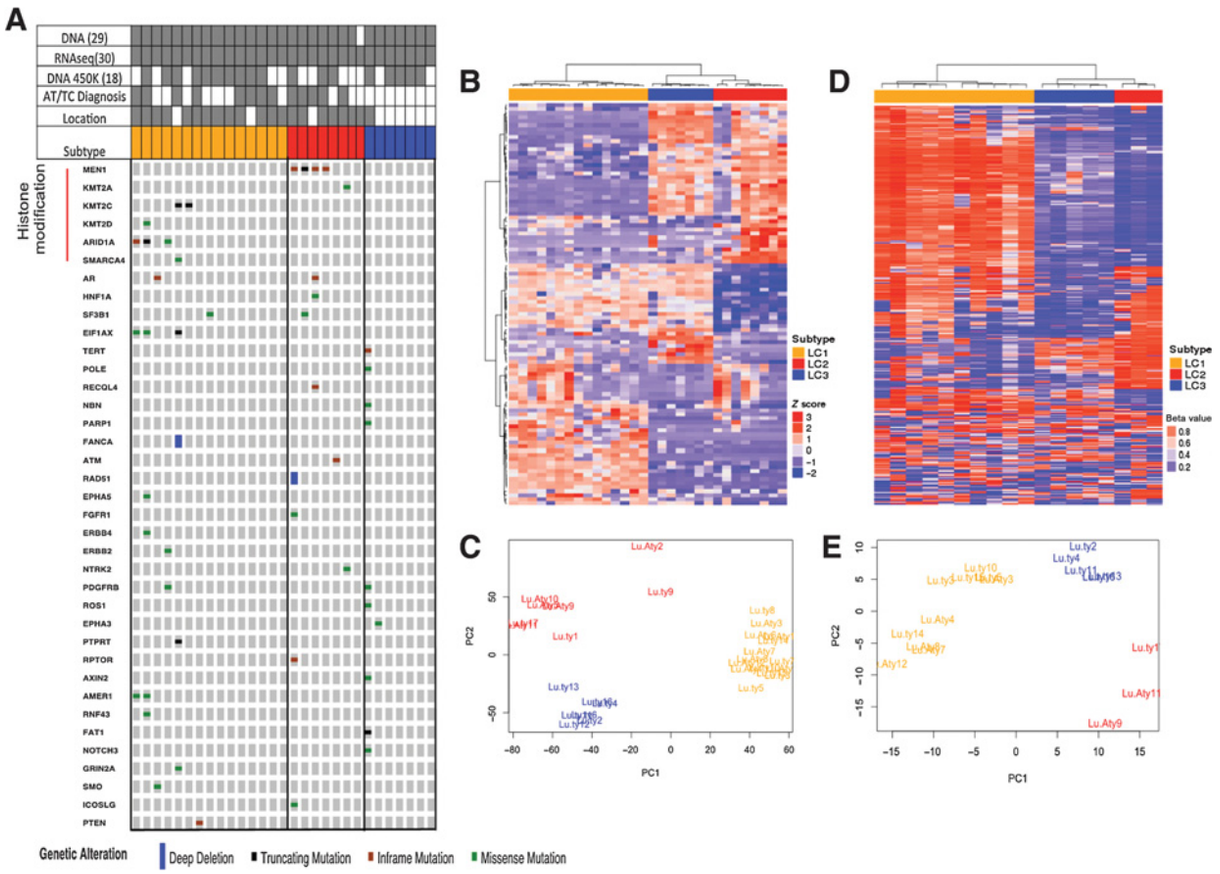


Figure 1.

Three novel molecular subtypes of LCs with mutational, gene expression, and DNA CpG methylation with distinct clinical features. **A,** Mutated genes in LCs on a 354-cancer gene panel. Samples summary from DNA ($n = 29$), methylation ($n = 18$), RNA-seq ($n = 30$), carcinoids samples (atypical ($n = 13$; gray), and typical ($n = 17$; white) and specimen location [endobronchial, white ($n = 9$); peripheral, gray ($n = 21$)]. Samples are grouped according to their gene expression and DNA methylation pattern. Orange, subtype 1 (LC1); red, subtype 2 (LC2; blue, subtype 3 (LC3). Column represents sample and row represents gene name. Gene expression ($n = 30$) and DNA CpG methylation ($n = 18$) analysis revealed three LC subtypes using unsupervised clustering and principal component analysis. **B**, Heatmap of unsupervised clustering of top 100 variably expressed gene across all samples. **C,** Principal component analysis of top 3,000 variably expressed genes. **D,** Heatmap of unsupervised clustering of top 500 variable methylated CpG probes. **E,** Principal component analysis of top 3,000 variable methylated CpG probes.

GSE118132 for 450 K methylation). We do not impose any restrictions on data availability.

Results

Patient cohort, clinical annotations, and mutational profile of LCs

We analyzed 30 randomly selected and histologically confirmed, well-differentiated LCs (17 TCs and 13 ACs) comprising the discovery dataset. Most specimens were from pulmonary lobectomy with lymph node detection. Tumor locations, that is, peripheral versus central (endobronchial), were assessed by combination of radiographic reveal and pathologic observations. Fifty-four percent (7/13) of ACs had either lymph node or distant metastasis, whereas 6% (1/17) of TCs had local lymph node metastasis. The 5-year disease-specific survival was 89% and 55% for TC and AC, respectively. Clinical information and features are presented in Supplementary File S1. In addition, a TMA containing 173 cases of LCs had been prepared previously (26) and used for study of clinical correlates.

We performed targeted sequencing of a 354-cancer gene panel (MSK-IMPACT; ref. 11) on 29 LCs from the discovery dataset. The mutated genes were enriched for those implicated in covalent histone modification/chromatin remodeling and found in 10 samples [*MEN1 (13.8%)*, *ARID1A (10%)*, *KMT2A* (3%), *KMT2C* (7%), *KMT2D* (3%), and *SMARCA4* (3%)] reproducing the results from a previous study (9). We also found mutations in DNA repair pathways (17.2% of samples) (Supplementary File S2; Fig. 1A). Mutations were not detected in the 354-cancer gene panel for 13 LC samples. Mutations in *MEN1*, the most frequently mutated gene, were found in 4 samples (4 ACs) and 4 of these mutations had variant allele frequencies higher than 70% indicating LOH (Supplementary Fig. S1). One sample (Lu-Aty9) has 2 *MEN1* mutations (an in-frame deletion and a missense substitution), a few bases apart on the same copy of *MEN1* (Supplementary File S2). The *ARID1A* gene is mutated in 3 samples with LOH occurring in one of the 3 samples. Using variant allele frequencies and LOH status of *MEN1* and *ARID1A*, we found median tumor purity to be 91% (Supplemental File S1), consistent with our pathology-based estimates (Supplementary Table S1). In

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

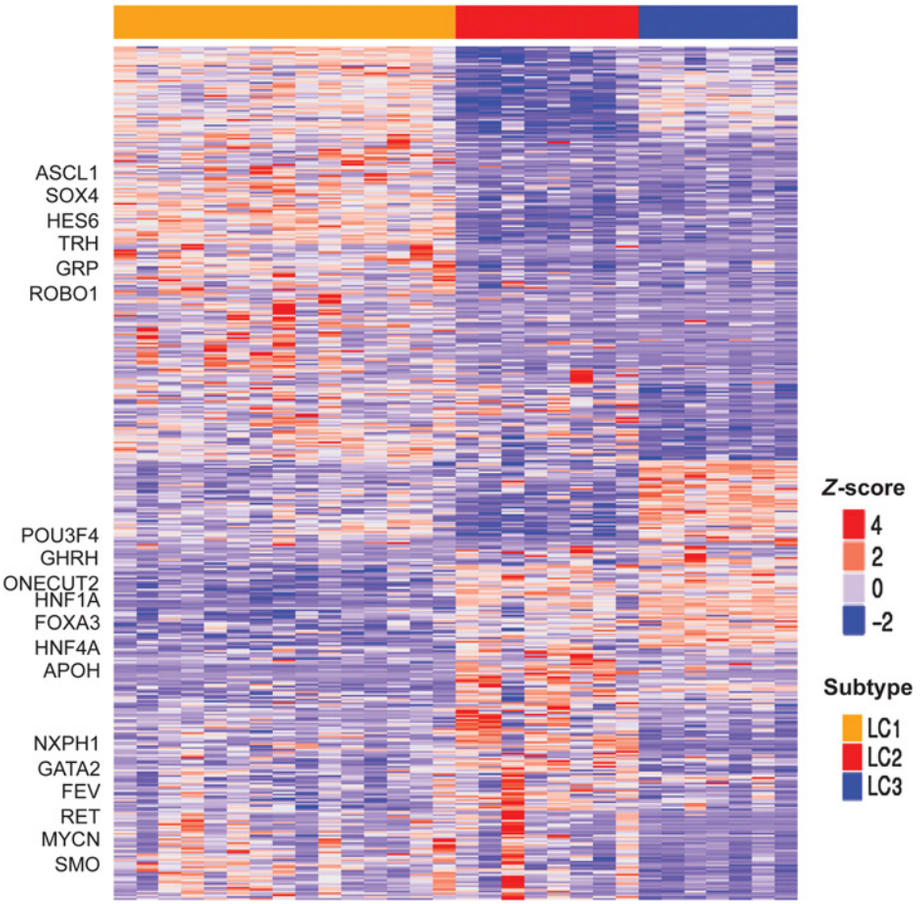


Figure 2.
Heatmap of differentially expressed genes between LC subtypes. Supervised analysis on LC subtypes reveals differential expression of transcription factor and neuropeptide (some are highlighted on the left side of the heatmap). Heatmap expression level is in *z*-score.

addition to *MEN1*, other genes encoding SET1/MLL complex proteins are also found mutated [*KMT2A* (3%), *KMT2C* (7%), *KMT2D* (3%)]. One sample (Lu-ty4) has the highest number of mutations (9) including mutations in *POLE* (DNA polymerase B domain: V1016M), *ROS1*, *FAT1*, *NBN*, *PARP1*, and *TERT* (in-frame deletion close to Telomerase RBD). We found homozygous deletions only in the *FANCA* and *RAD51* genes in 2 different ACs. The most recurrent CNV are single copy deletions in *FANCA* (17%), *FAT1* (10%), *MEN1* (7%), *ATM* (17%), *SDHD* (17%), and *CHEK1* (17%), many of which reside on chr11q. We did not observe changes in the transcription levels of these genes with hemizygous deletions in comparison to wild-type samples. There are 18 samples (4 ACs and 14 TCs) with normal karyotype, 6 samples (4 ACs and 2 TCs) with nearly normal karyotype (aneuploid for only one or 2 different chromosomes), and 6 samples (5 ACs and 1 TCs) with aneuploidy in more than 2 different chromosomes in our dataset (Supplementary File S2). We did not find any known pathogenic germline mutations in the panel of cancer-associated genes in our samples. *TP53* and *RB1* genes were not mutated in this cohort, unlike high-grade lung NETs and SCLC.

Transcriptome and methylome profiles reveal three distinct subtypes

We performed RNA-seq on 30 LCs (13 ACs and 17 TCs) and DNA methylation analysis on 18 LCs (12 of the 30 samples did not have sufficient material for analysis) from the discovery dataset. Unsupervised clustering and principal component analysis on the top 3,000 variable genes showed 3 distinct clusters (Fig. 1B and C). These clusters are robust when different numbers of top variable genes were used for clustering (Supplementary Fig. S2). We named these subtypes LC1, LC2, and LC3. Heatmap of Pearson correlation on top 3,000 variable genes shows 3 blocks representing the 3 evident subtypes (Supplementary Fig. S3). The top 100 variable genes (Supplementary Fig. S4) across all LCs show enrichment for gene ontologies related to hormonal secretions, endogenous stimulus, wound healing, and developmental processes (Supplementary Table S2). Gene expression analysis revealed greater similarity between LC2 and LC3 as compared with LC1 (Supplementary Figs. S3 and S4).

We investigated the DNA methylation profiles of LCs ($n = 18$), using the Illumina 450K microarray. Unsupervised clustering and PCA of the top 3,000 variable CpG sites revealed 3 distinct subtypes in complete agreement with the gene expression based subtypes (Fig. 1D and E). Consistent with gene expression, we also observed greater similarity of DNA methylation levels for LC2 and LC3 subtypes when compared with LC1. The 3 grouping of subtypes was robust and reproducible using different numbers of top variable CpG sites (Supplementary Fig. S5). Total genome-wide DNA methylation level is not different between the 3 subtypes (Supplementary Table S3).

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

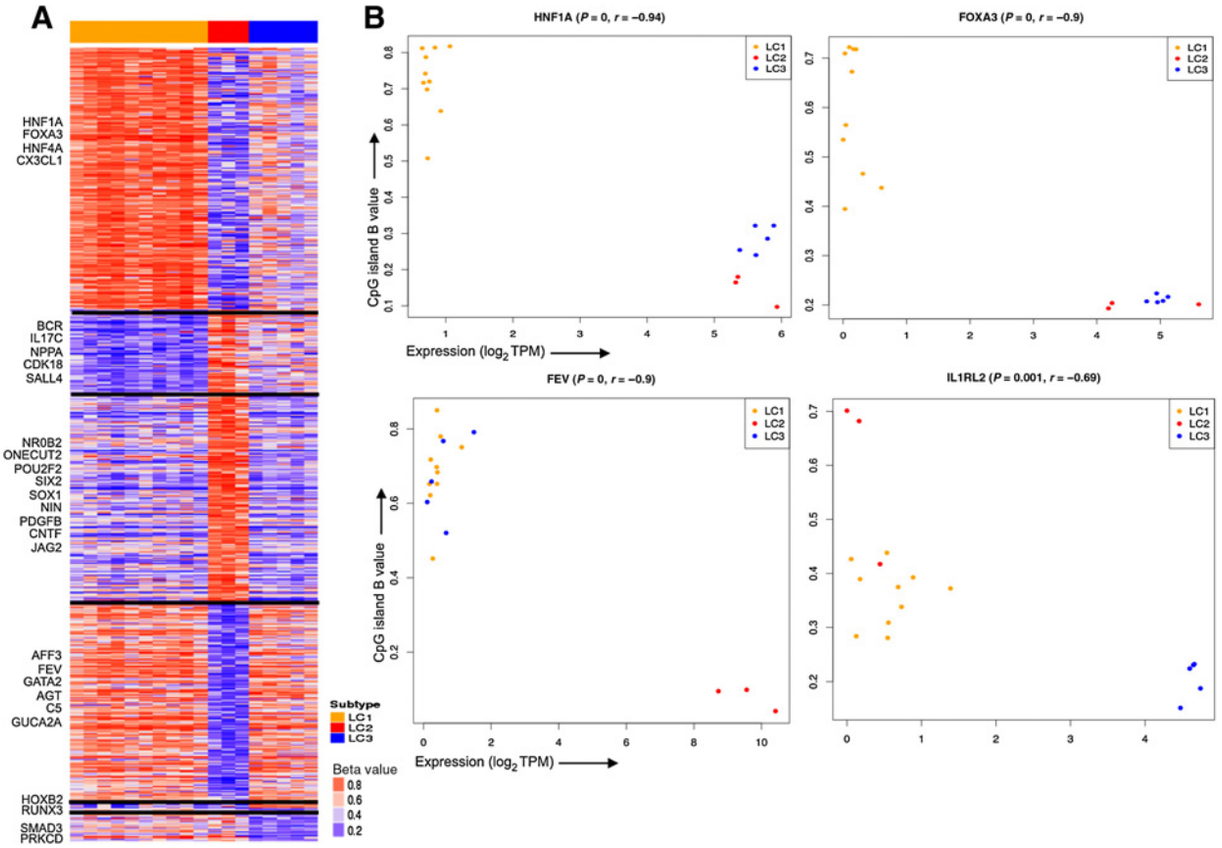


Figure 3.

Subtype-specific molecular characterization of gene expression and DNA methylation profiles. **A,** Heatmap of differentially methylated CpG sites (probes from TSS1500, TSS200, and first exon) of genes among the three LC subtypes. Some genes with altered gene expression and CpG sites are highlighted on the left of heatmap. Dark black line, subtype-specific blocks. **B,** Anticorrelation of gene expression and respective CpG island methylation (18 matched samples) for *HNF1A, FOXA3, FEV*, and *ILRL2* across three subtypes. Each plot represents gene expression on *x*-axis and average CpG island β value on *y*-axis along with Pearson correlation (*r*) and *P* value (*P*) are on top of the plot.

Subtype-specific molecular characterization of LCs

We investigated gene expression and CpG DNA methylation profiles to determine subtype-specific molecular alterations (see Materials and Methods section). Genes upregulated in LC2 and LC3 as compared with LC1 are enriched for having the transcription factor motifs for *HNF1* (FDR *q*-value <0.001) and *HNF4* (FDR *q*-value <0.001; Fig. 2; Supplementary File 3). This is in agreement with the observed high gene expression and DNA hypomethylation of *HNF1A*, *FOXA3*, and *HNF4A* in LC2 and LC3 as compared with LC1 (Fig. 3A and B; Supplementary Fig. S6). In addition, many of the most highly expressed genes (*APOH*, *GC*, *HAO1*, *G6PC*, *TM4SF4*, *PKLR*, *UGT2B17*, *CDH1*, and *SERPINA1/2/6*) in LC2 and LC3 are targets of these hepatocyte nuclear factors (Fig. 2). Cancer hallmark gene set enrichment analysis shows complement and coagulation, xenobiotic, retinol and bile acid metabolism to be significantly upregulated in LC2 and LC3 as compared with LC1, a gene signature also found in subset of pancreatic neuroendocrine tumors (Supplementary File 3; ref. 14). However, we also identified TFs that are differentially expressed between LC2 and LC3 (*FEV* and *POU3F4* are more highly expressed in LC2 and LC3, respectively; Fig. 2; Supplementary File S3). *MEN1* gene is required for regulation of several members of the HOX gene family (27). Indeed, the LC2 subtype, which included all of the *MEN1* mutant samples, has low expression of *HOXB2/3/4/5/6* genes as compared with LC1 and LC3 (Supplementary Fig. S7).

We integrated subtype-specific CpG DNA methylation (see Materials and Methods section) with gene expression by focusing on CpG sites between 1,500 bps and 200 bps upstream to the transcription start site (TSS) and in the first exon, which have been shown to inversely correlate with gene expression (28). Fig. 3A shows subtype-specific differentially methylated CpG probes (DMP) and their inverse correlation with neighboring gene expression. We found 75 genes with expression to be significantly anticorrelated with respective CpG island methylation level (FDR *P*-value < 0.01; Supplementary File S4). *HNF1A* and *FOXA3* are hypermethylated and low expressed in LC1. *FEV*, *GATA2*, and *PROCR* are hypomethylated and highly expressed in LC2. *SOX1* is hypermethylated and low expressed in LC2. *SIX2*, *ONECUT2*, and *IL1RL2* are hypomethylated and highly expressed in LC3 (Fig. 3B). Many of these observations suggest further mechanistic studies but there are currently no appropriate LC cell lines or animal models available.

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

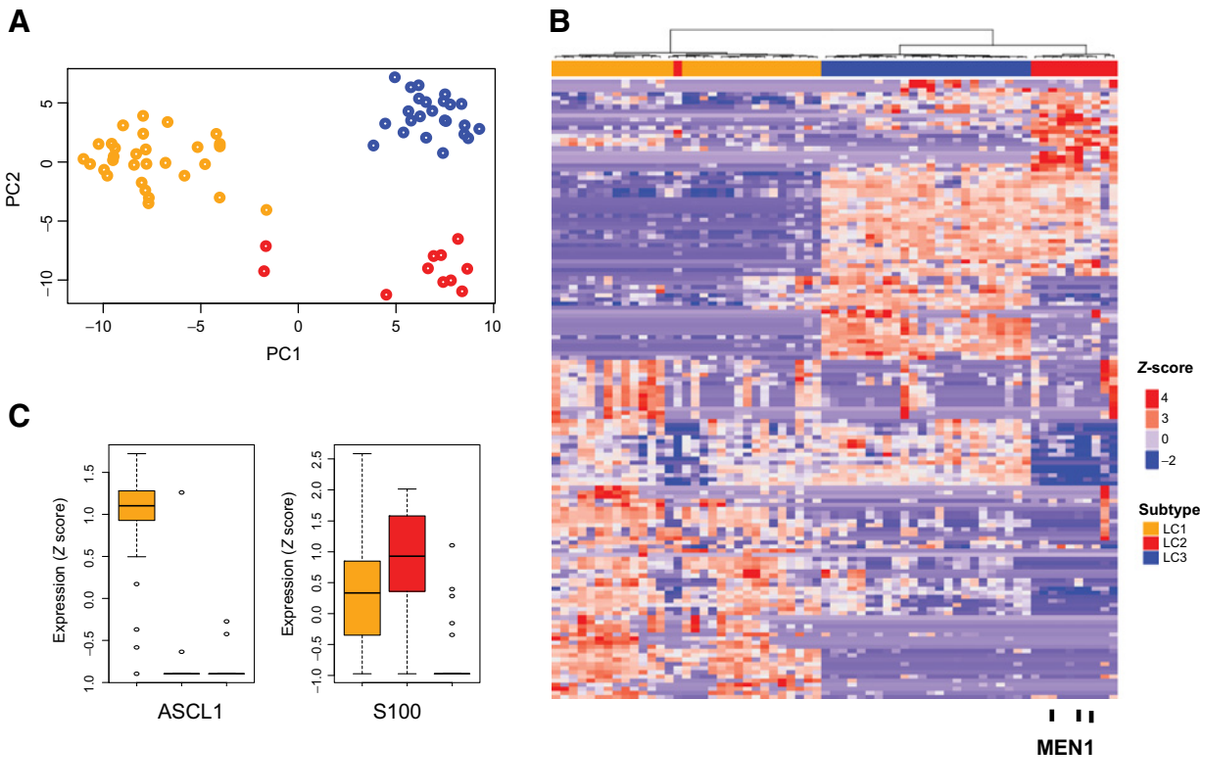
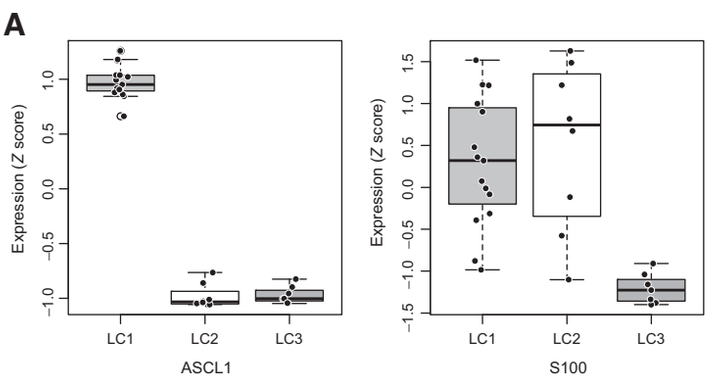


Figure 4.

Validation of novel classification of LC on an independent collection of LCs from Fernandez-Cuesta and colleagues (9). **A** and **B,** Principal component analysis and heatmap of hierarchical clustering of gene expression of LCs from Fernandez-Cuesta and colleagues (9) using our top 100 variable gene set signature shows three distinct subtypes LC1 (orange), LC2 (red), and LC3 (blue). Black sticks represent samples with *MEN1* mutations and they are all found in subtype LC2. **C,** Boxplot of *ASCL1* and *S100* gene expression from Fernandez-Cuesta and colleagues (9) is consistent with LC subtypes. Centerline, median; bounds of box, the first and third quartiles; and upper and lower whisker is defined to be 1.5 × interquartile range more than the third and first quartile.

Independent validation of LC classification

We validated our novel classification and gene expression biomarkers using published LC data from Fernandez-Cuesta and colleagues (9), which include genome/exome and RNA sequencing of 65 samples (56 TCs, 6 ACs and 3 carcinoids). Using our gene signatures derived from the top 100 most variable genes (Supplementary File S5) across LCs, we found 3 distinct subtypes using unsupervised clustering and PCA that are consistent with the subtypes identified from our data (Fig. 4A and B; Supplementary Fig. S8). Moreover, all *MEN1* mutated LCs are found exclusively in LC2 (Fig. 4B). In addition, we found *HNF1A* and *FOXA3* are more highly expressed in LC2 and LC3 as compared with LC1 whereas



B

Subtype	*ASCL1* (a-MASH1)	*S100*
	Positive	Positive
LC 1 (*n* = 11)	100%	0
LC 2 (*n* = 5)	0	100%
LC 3 (*n* = 4)	0	0

Figure 5.

Gene expression and immunohistochemistry for ASCL1 and S100 biomarker genes. **A,** Boxplot of *ASCL1* and *S100* gene expression. Centerline, median; bounds of box, the first and third quartiles; and top and bottom whisker is defined to be 1.5 × interquartile range more than the third and first quartile. **B,** IHC staining results for ASCL1 and S100 in LC samples: LC1 ($n = 11$), LC2 ($n = 5$), and LC3 ($n = 4$). Supplementary Table S4 has IHC results for all samples.

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

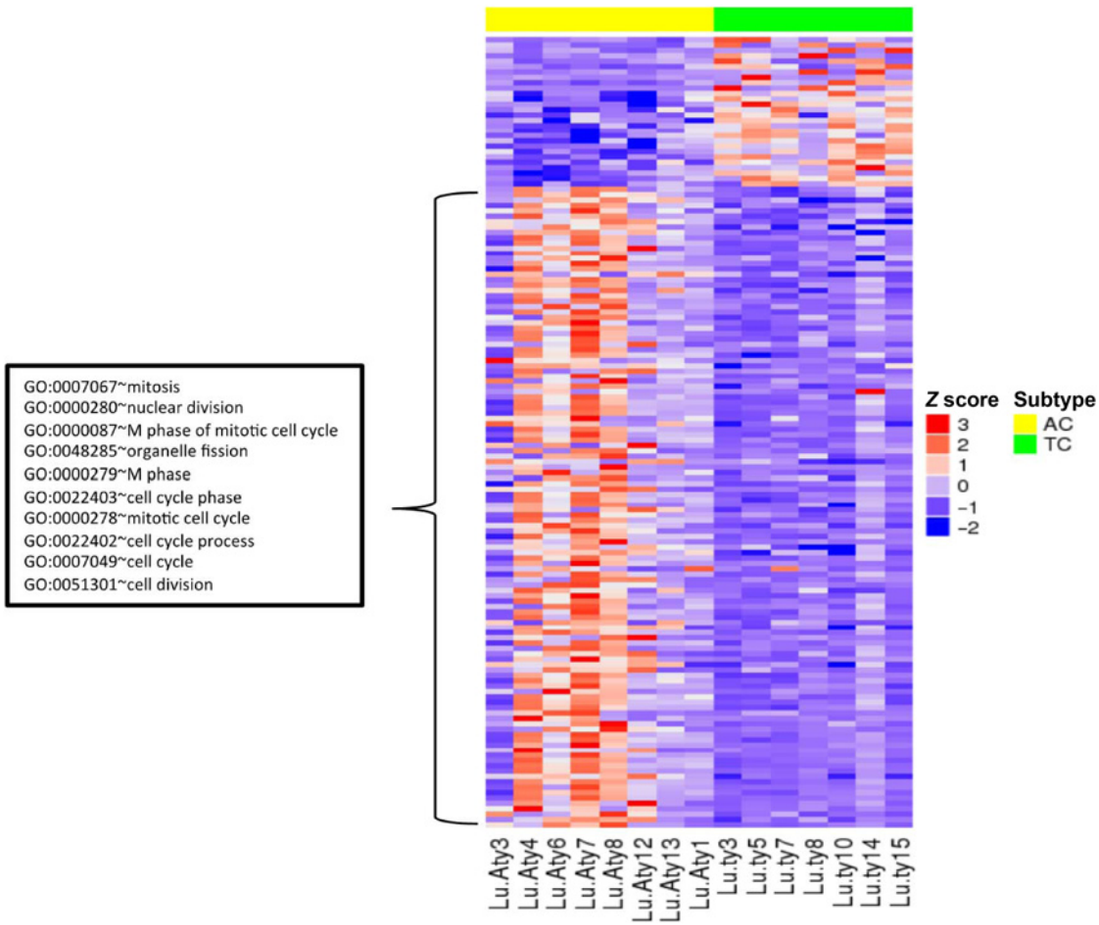


Figure 6.
Heatmap of differentially expressed genes between ACs and TCs within LC1 subtype. Upregulated genes in ACs (of LC1) are significantly enriched for genes involved in cell cycle/mitosis.

FEV is highly expressed only in LC2 consistent with our data (Supplementary Fig. S9).

ASCL1 and *S100* are novel biomarkers for LC subtypes

We selected genes with distinct subtype-specific expression to test for use as biomarkers. *ASCL1* encodes a transcription factor that plays a role in neuronal differentiation and proliferation (29), neuroepithelial bodies formation (30), and is a lineage-specific oncogene for high-grade neuroendocrine lung cancer (31). *ASCL1* is significantly highly expressed in LC1 along with its transcriptional targets (Figs. 4C and 5A; Supplementary Fig. S10). *S100*, a family of proteins containing 2 EF-hand calcium-binding motifs, is implicated in tumor progression and poor prognosis (32). Its gene expression levels are significantly higher in subtype LC2 (Fig. 5A). We performed IHC staining of *ASCL1* and *S100* to use as biomarkers. *ASCL1* stained positively only for LC1 samples ($n = 11$) and *S100* stained positively only for LC2 samples ($n = 5$) (Fig. 5B). Both of these genes stained negatively for LC3 samples ($n = 4$; Supplementary Table S4).

Additionally, we performed *ASCL1* and *S100* IHC staining on a panel of 173 LCs TMA (Supplementary File S6). *ASCL1* positive and *S100* negative samples ($n = 54$) were designated LC1. *ASCL1* negative and *S100* positive samples ($n = 15$) were designated LC2. *ASCL1* negative and *S100* negative samples ($n = 71$) were designated LC3. Fifteen percent of the TMA samples stained positive for *ASCL1* and *S100*, which are not represented in our discovery dataset.

Cell cycle and mitotic genes are highly expressed in ACs of LC1

Pathologically, ACs are more aggressive and have a higher mitotic index in comparison to TCs. To find the gene signature responsible for these features of ACs, we compared ACs ($n = 13$) and TCs ($n = 17$) from our 30 LC cohort. Surprisingly, we did not find cell cycle or mitosis-related genes to be differentially expressed. We then controlled for LC subtypes and compared ACs ($n = 8$) and TCs ($n = 7$) from subtype LC1 and found differentially expressed genes (Fig. 6) were then enriched for mitotic and cell cycle related pathways with high expression in the ACs (Supplementary File 7). Of the 8 AC tumors, the 3 with highest gene expression signature for mitotic/cell-cycle pathway have metastases or recurrences whereas only 1 of the 5 with lower gene expression signature have recurrence or metastases (Fig. 6). We also observed high aneuploidy in the ACs with high gene expression signature of mitotic/cell-cycle pathway. We performed

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

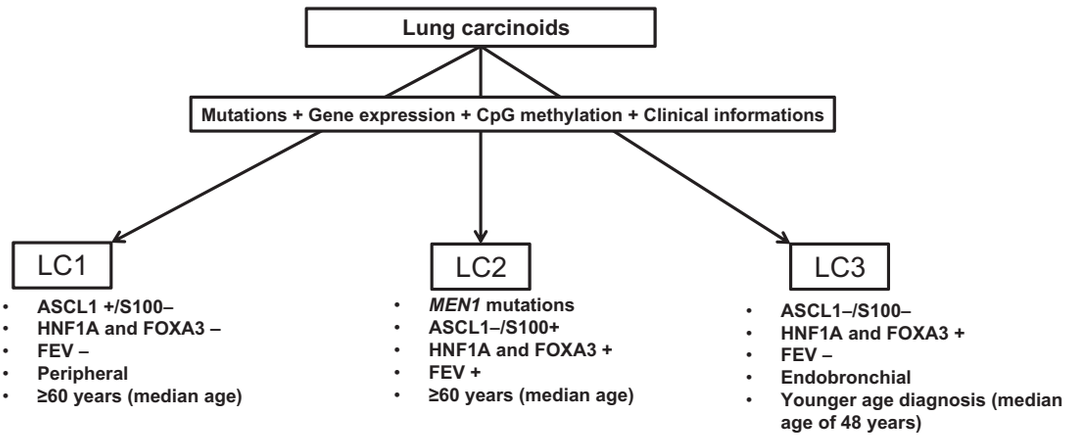


Figure 7.
Schematic representation of novel molecular subtypes of LCs with the principal genomic and clinical characteristics.

the same analysis for ACs and TCs from LC2 and did not find any significant gene signatures, which may be due to the small sample size of LC2.

LC subtypes have distinct clinical phenotypes

The 3 subtypes of LCs have distinct clinical phenotypes. Subtype LC1 is enriched for peripheral lung (P-value <0.003 in discovery dataset; P-value <0.002 in TMA), whereas subtype LC3 is found predominately at endobronchial lung (P-value < 0.054 in discovery dataset; P-value <3.8e−5 in TMA; Fig. 1A, box; Supplementary File S1). Subtype LC3 has significantly younger age of diagnosis (median age of 33, 44.5, and 48 years in discovery, Fernandez-Cuesta and colleagues (9), and TMA datasets, respectively) than LC1 (median age of 67, 66, and 60 years, respectively) and LC2 (median age of 62.5, 57, and 65 years, respectively; Supplementary Fig. S1A and S1B). LC1 subtype was enriched for female patients (6.5-fold, P-value < 0.007 in discovery dataset; 3.9-fold, P-value < 1.4e−5 in TMA) but not for LC2 or LC3 (Supplementary File S1).

Discussion

We identified 3 novel molecular subtypes of LCs with distinct clinical features using gene expression, DNA methylation, and mutational profiles (Fig. 7). Integrative analysis of gene expression and DNA methylation identified subtype-specific transcriptional profiles of key differentiation transcription factors (*ASCL1*, *HNF1A*, *FOXA3*) and their downstream target genes. Mutational analysis revealed recurrent mutations in chromatin remodeling genes found in all subtypes of LCs with exception of *MEN1* mutations occurring only in subtype LC2 tumors. Importantly, we found mutations in DNA repair genes in 17% of our LC samples. Subtype LC3 has younger age of diagnosis and is predominantly endobronchial, whereas subtypes LC1 are predominantly found in peripheral regions of the lung. These findings may argue that subtypes of LC potentially originate from different neuroendocrine cell lineages, although the lack of available cell-type–specific gene markers prevents a definitive validation beyond speculation at this time. Nevertheless, we believe that a more comprehensive single-cell approach can uncover lung NE cell-type–specific gene signatures and reveal the cells of origin for the LC subtypes. The younger age of diagnosis for LC3 by 15 to 20 years as compared with LC1 and LC2 may be due to earlier diagnosis from the clinically symptomatic tumors located in the central lung as compared with asymptomatic tumors at the peripheral lung or suggest a possible distinct pathogenesis predisposing to LC3, including germline mutations. However, we were not able to detect any pathogenic germline mutations in the panel of cancer-associated genes used in the MSK IMPACT testing for any of the LCs. Our classification and gene expression biomarkers were validated in 65 additional LC samples from Fernandez-Cuesta and colleagues (9). Using our subtype classification, we found gene signature for cell cycle and mitotic processes activated in ACs as compared with TCs of the LC1 subtype and those ACs with the high gene signature have worst outcome (Fig. 6; Supplementary File 7). This gene signature will need to be reproduced with a larger sample size but may potentially serve as a diagnostic and prognostic biomarker to differentiate malignant from more benign ACs from subtype LC1. This gene signature is specific to LC1 and would not have been found from comparing ACs to TCs from all LCs demonstrating the need to study distinct subtypes individually.

Our molecular classification introduces 3 subtypes of LCs with distinct clinical phenotypes. This can refine and complement the WHO classification of LCs into typical or ACs and help diagnose ambiguous cases of LCs from the more malignant LCNEC and SCLC. In addition, the stratification of LCs into distinct molecular subtypes will help with future study of their tumorigenesis, prognosis, and treatment options.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: S.V. Laddha, W.D. Travis, L.H. Tang, C.S. Chan
Development of methodology: S.V. Laddha, J.T. Poirier, C.S. Chan
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.V. Laddha, K. Robzyk, B.R. Untch, W.D. Travis, L.H. Tang

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.V. Laddha, B.R. Untch, H. Ke, W.D. Travis, L.H. Tang, C.S. Chan
Writing, review, and/or revision of the manuscript: S.V. Laddha, E.M. da Silva, K. Robzyk, B.R. Untch, N. Rekhtman, W.D. Travis, L.H. Tang, C.S. Chan
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S.V. Laddha, E.M. da Silva, K. Robzyk
Study supervision: S.V. Laddha, C.S. Chan

Acknowledgments

We would like to thank Starr Cancer Consortium Grant (L.H. Tang), Raymond and Beverly Sackler Foundation (L.H. Tang), Caring for Carcinoid Foundation (L.H. Tang), Mushett Family Foundation (L.H. Tang), MSKCC Support Grant/Core Grant (P30 CA008748), and Stand Up To Cancer (C.S. Chan). This material is based upon work supported in part by the National Science Foundation under Grant No. 1546101. This research was also supported by the Biomedical Informatics shared resource of Rutgers Cancer Institute of New Jersey (P30CA072720). Computational resources were provided by the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey, under the National Institutes of Health Grant (S10OD012346).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 22, 2019; revised June 3, 2019; accepted July 9, 2019; published first July 12, 2019.

References

1. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. J Thorac Oncol 2015;10:1243–60.
2. Caplin ME, Baudin E, Ferolla P, Filosso P, Garcia-Yuste M, Lim E, et al. Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. Ann Oncol 2015;26:1604–20.
3. Travis WD, Rush W, Flieder DB, Falk R, Fleming MV, Gal AA, et al. Survival analysis of 200 pulmonary neuroendocrine tumors with clarification of criteria for atypical carcinoid and its separation from typical carcinoid. Am J Surg Pathol 1998;22:934–44.
4. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. Acta Neuropathol 2010;120:297–304.
5. Swarts DR, van Suylen RJ, den Bakker MA, van Oosterhout MF, Thunnissen FB, Volante M, et al. Interobserver variability for the WHO classification of pulmonary carcinoids. Am J Surg Pathol 2014;38:1429–36.
6. Pelosi G, Papotti M, Rindi G, Scarpa A. Unraveling tumor grading and genomic landscape in lung neuroendocrine tumors. Endocr Pathol 2014;25:151–64.
7. Volante M, Gatti G, Papotti M. Classification of lung neuroendocrine tumors: lights and shadows. Endocrine 2015;50:315–9.
8. Pelosi G, Rodriguez J, Viale G, Rosai J. Typical and atypical pulmonary carcinoid tumor overdiagnosed as small-cell carcinoma on biopsy specimens: a major pitfall in the management of lung cancer patients. Am J Surg Pathol 2005;29:179–87.
9. Fernandez-Cuesta L, Peifer M, Lu X, Sun R, Ozretic L, Seidal D, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nat Commun 2014;5:3518.
10. Vollbrecht C, Werner R, Walter RF, Christoph DC, Heukamp LC, Peifer M, et al. Mutational analysis of pulmonary tumours with neuroendocrine features using targeted massive parallel sequencing: a comparison of a neglected tumour group. Br J Cancer 2015;113:1704–11.
11. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015;17:251–64.
12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–9.
14. Chan CS, Laddha SV, Lewis PW, Koletsky MS, Robzyk K, Da Silva E, et al. ATRX, DAXX or MEN1 mutant pancreatic neuroendocrine tumors are a distinct alpha-cell signature subgroup. Nat Commun 2018;9:4158.
15. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13.
16. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006;38:500–1.
17. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.
18. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
20. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
22. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics 2011;12:322.
23. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics 2014;30:428–30.
24. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol 2014;15:503.
25. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, et al. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res 2013;41:e117.
26. Rekhtman N, Pietanza CM, Sabari J, Montecalvo J, Wang H, Habeeb O, et al. Pulmonary large cell neuroendocrine carcinoma with adenocarcinoma-like features: napsin A expression and genomic alterations. Mod Pathol 2018;31:111–21.
27. Yokoyama A, Somervaille TC, Smith KS, Rozenblatt-Rosen O, Meyerson M, Cleary ML. The menin tumor suppressor protein is an essential oncogenic cofactor for MLL-associated leukemogenesis. Cell 2005;123:207–18.
28. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One 2011;6:e14524.
29. Castro DS, Martynoga B, Parras C, Ramesh V, Pacary E, Johnston C, et al. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes Dev 2011;25:930–45.
30. Guha A, Vasconcelos M, Cai Y, Yoneda M, Hinds A, Qian J, et al. Neuroepithelial body microenvironment is a niche for a distinct subset of Clara-like precursors in the developing airways. Proc Natl Acad Sci U S A 2012;109:12592–7.
31. Borromeo MD, Savage TK, Kollipara RK, He M, Augustyn A, Osborne JK, et al. ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. Cell Rep 2016;16:1259–72.
32. Chen H, Xu C, Jin Q, Liu Z. S100 protein family in human cancer. Am J Cancer Res 2014;4:89–115.

Published OnlineFirst July 12, 2019; DOI: 10.1158/0008-5472.CAN-19-0214

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Integrative Genomic Characterization Identifies Molecular Subtypes of Lung Carcinoids

Saurabh V. Laddha, Edaise M. da Silva, Kenneth Robzyk, et al.

Cancer Res 2019;79:4339-4347. Published OnlineFirst July 12, 2019.

Lung Cancer 145 (2020) 85–94

Contents lists available at ScienceDirect

Lung Cancer

journal homepage: www.elsevier.com/locate/lungcan

Whole-exome and RNA sequencing of pulmonary carcinoid reveals chromosomal rearrangements associated with recurrence

Akihiko Miyanaga[a,b,*], Mari Masuda[a], Noriko Motoi[c], Koji Tsuta[c], Yuka Nakamura[a], Nobuhiko Nishijima[a,b], Shun-ichi Watanabe[d], Hisao Asamura[d], Akihiko Tsuchida[e], Masahiro Seike[b], Akihiko Gemma[b], Tesshi Yamada[a,e]

[a] Division of Chemotherapy and Clinical Research, National Cancer Center Research Institute, Tokyo, Japan
[b] Department of Pulmonary Medicine and Oncology, Graduate School of Medicine, Nippon Medical School, Tokyo, Japan
[c] Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Tokyo, Japan
[d] Division of Thoracic Surgery, National Cancer Center Hospital, Tokyo, Japan
[e] Department of Gastrointestinal and Pediatric Surgery, Tokyo Medical University, Tokyo, Japan

ARTICLE INFO

Keywords:
Pulmonary carcinoid
Next-generation sequencing
Chromosomal rearrangement
MUC gene family
Postsurgical recurrence

ABSTRACT

Introduction: The majority of pulmonary carcinoid (PC) tumors can be cured by surgical resection alone, but a significant proportion of patients experience recurrence. As PC is insensitive to conventional chemotherapy, further clarification of the molecular mechanisms of metastasis is needed in order to develop targeted therapeutics.

Methods: We performed comprehensive whole-exome sequencing (WES) of primary tumors and corresponding normal lung tissues from 14 PC patients (including 4 patients who developed postsurgical distant metastasis) and RNA sequencing of primary tumors from 6 PC patients (including 4 patients who developed postsurgical distant metastasis). Exon array-based gene expression analysis was performed in 25 cases of PC.

Results: We identified a total of 139 alterations in 136 genes. *MUC6* and *SPTA1* were recurrently mutated at a frequency of 21% (3/14) and 14% (2/14), respectively. Mucin protein family genes including *MUC2*, *MUC4* and *MUC6* were mutated in a mutually exclusive manner in 36% (5/14). Pathway analysis of the mutated genes revealed enrichment of genes involved in mitogen-activated protein kinase (MAPK) signaling, regulation of the actin cytoskeleton and focal adhesion, and transforming growth factor (TGF)-β signaling. RNA sequencing revealed a total of 8 novel fusion transcripts including one derived from a chromosomal translocation between the *TRIB2* and *PRKCE* genes. All of the 8 fusion genes were detected in primary PCs that had developed metastasis after surgical resection. We identified 14 genes (*DENND1B*, *GRID1*, *CLMN*, *DENND1B*, *NRP1*, *SEL1L3*, *C5orf13*, *TNFRSF21*, *TES*, *STK39*, *MTHFD2*, *OPN3*, *MET*, and *HIST1H3C*) up-regulated in 5 PCs that had relapsed after surgical resection.

Conclusions: In this study we identified novel somatic mutations and chromosomal rearrangements in PC by examining clinically aggressive cases that had developed postsurgical metastasis. It will be essential to validate the clinical significance of these genetic changes in a larger independent patient cohort.

1. Introduction

Neuroendocrine (NE) tumors of the lung comprise four distinct histologic types: typical carcinoid (TC), atypical carcinoid (AC), large cell neuroendocrine carcinoma (LCNEC), and small cell lung carcinoma (SCLC), as listed in the 2015 World Health Organization (WHO) classification of tumors of the lung, pleura, thymus and heart [1]. Although these four tumor types are grouped into the same category, PC (TC and

Abbreviations: **AC**, atypical carcinoid; **CIN**, chromosomal instability; **ESS**, endometrial stromal sarcoma; **FISH**, fluorescence *in situ* hybridization; **KEGG**, Kyoto Encyclopedia of Genes and Genomes; **LCNEC**, large cell neuroendocrine carcinoma; **MAPK**, mitogen-activated protein kinase; **NCC**, National Cancer Center; **NE**, neuroendocrine; **NGS**, next-generation sequencing/sequencer; **PC**, pulmonary carcinoid; **PRKCE**, protein kinase C epsilon; **SCLC**, small cell lung carcinoma; **SNP**, single nucleotide polymorphism; **TC**, typical carcinoid; **TGF-β**, transforming growth factor-β; **WES**, whole-exome sequencing; **WHO**, World Health Organization

* Corresponding author at: Department of Pulmonary Medicine and Oncology, Graduate School of Medicine, Nippon Medical School, 1-1-5 Sendagi, Bunkyo-ku, Tokyo 113-8602, Japan.
E-mail address: a-miyanaga@nms.ac.jp (A. Miyanaga).

https://doi.org/10.1016/j.lungcan.2020.03.027
Received 31 December 2019; Received in revised form 24 March 2020; Accepted 27 March 2020

AC) is not an early progenitor lesion of LCNEC or SCLC [2]. PC has distinct epidemiologic, clinical, histological, and genetic characteristics, and its incidence is not associated with cigarette smoking. Patients with PC are significantly younger and have a better prognosis than those with LCNEC or SCLC [3].

In the WHO classification, TC is defined as a low-grade neuroendocrine tumor lacking necrosis and having less than 2 mitoses per 2 mm^2. AC is defined as an intermediate-grade neuroendocrine tumor displaying 2–10 mitoses per 2 mm^2 and/or the presence of necrosis. Necrosis is usually absent, but if present it tends to be focal or punctate [1]. In our large retrospective study, the 5-year survival rate of patients with completely resected TC was 96% [4]. AC is prone to locoregional lymph node and distant metastasis and has a 5-year survival rate of 78% [4]. PC is generally insensitive to chemotherapy and radiotherapy, and no effective treatment has been established for PC patients with systemic metastasis and those who develop postsurgical metastasis [5–7]. PC patients with distant metastasis at the time of diagnosis reportedly have a 5-year survival rate of 14–25% [8].

The emergence of so-called next-generation sequencing (NGS) technologies has enabled rapid genome-wide surveys of oncogenic and tumor suppressive signaling molecules in various cancers [9–11] and accelerated understanding of cancer biology and the development of novel diagnostics and therapeutics. Recent large-scale sequencing studies have revealed that PC has a large variety of genetic abnormalities [2,12,13], but it is still unclear which genetic alterations or pathways are key players the development and progression of PC.

In the present study, we performed a comprehensive whole exon (exome) and RNA sequencing analysis of PC and revealed recurrent mutations in the mucin protein family genes. Here we report a novel association of chromosomal rearrangements with the postsurgical metastasis of PC.

2. Materials and methods

2.1. Tissues samples

The study protocol was reviewed and approved by the Institutional Review Board of the National Cancer Center (NCC) (Tokyo, Japan). Informed consent was obtained from 25 patients with PC (20 TC and 5 AC) who underwent lobectomy at the NCC Hospital (Tokyo, Japan) between 1998 and 2010 prior to specimen collection. Tissue samples were snap-frozen in liquid nitrogen within 30 min of surgical resection and stored by the National Center Biobank Network (NCBN). The histopathological diagnosis of PC was made by expert pathologists in accordance with the 2015 WHO classification.

2.2. DNA and RNA preparation, cDNA synthesis, and transcriptome analysis

Genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's protocol. Total RNA was extracted using TRIzol reagent (Invitrogen) or RNeasy Mini Kit columns (Qiagen) in accordance with the manufacturer's protocol. RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). All samples showed RNA Integrity Numbers of > 7.0. The quantity of genomic DNA and total RNA was determined using a NanoDrop 2000 Spectrophotometer (Thermo Scientific). rRNA reduction, first-round double-strand cDNA synthesis, cRNA synthesis, second-round single-strand (ss)-cDNA synthesis, ss-cDNA fragmentation, and labeling were performed in accordance with the Affymetrix GeneChip Whole-Transcript Sense Target–Labeling Assay manual. Affymetrix Human Exon 1.0 ST arrays were hybridized overnight with 5 μg biotin-labeled ss-cDNA.

2.3. Wes

Three micrograms of genomic DNA was used to construct libraries for sequencing. The quality of the constructed libraries was assessed using an Agilent 2100 Bioanalyzer. All the exon genes were captured using a SureSelect^{XT} Human All Exon kit v3 (Agilent) in accordance with the SureSelect^{XT} Target Enrichment for Illumina Paired-End Multiplexed Sequencing Protocol 1.1.1. Enriched libraries were sequenced using an Illumina Genome Analyzer IIx. Base calling was performed using the Illumina Pipeline (CASAVA v1.8) with default parameters. Adaptor trimming was performed using cutadapt v1.2.1 with the parameters "-O 9 -m 32″. Read cleaning was performed using the "fastq_quality_rimmer -l 32″ and "fastq_quality_filter -q 10 -p 95″ commands in FASTX-Toolkit v 0.0.13 followed by paired reads extraction using cmpfastq_pe. Cleaned reads were mapped on UCSC hg19 using the "bwa aln" and "bwa sampe" commands with default paramaters in BWA v0.5.9. Duplicate reads were marked using Picard MarkDuplicates. Realignment and base recalibration were performed using RealignerTargetCreator/IndelRealigner and CountCovariates in GATK v1.6, respectively. Somatic variant calling was performed using Strelka v1.0.14 and Virmid v1.1.1 with default parameters. Variants were annotated using SnpEff v3.6c.

2.4. RNA sequencing

Total RNA (14 − 36 μg) was used for the construction of libraries using the TruSeq RNA Sample Preparation Kit (v2, Illumina). Libraries were used to generate clustered flowcells on cBot using the TruSeq PE Cluster Kit v2. Paired-end sequencing (75-base) was performed on an Illumina Genome Analyzer IIx sequencer using a TruSeq SBS Kit v5. The Illumina software pipeline was used for processing of image data into raw sequencing data (SCS v2.9 and CASAVA v1.8.1). Sequence reads marked as "passed filtering" were used for selection of fusion genes using the deFuse program [14,15].

2.5. Dideoxynucleotide sequencing

Total RNA was reverse-transcribed into cDNA using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) with random priming. Genomic DNA or cDNA was amplified using EmeraldAmp PCR Master Mix (TaKaRa Bio) with relevant primer pairs (Supplementary Table S1) designed by Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/input.htm). The amplified PCR products were sequenced using the Big Dye Terminator Cycle Sequencing kit on an ABI 3100 genetic analyzer (Applied Biosystems). Sequence data were analyzed using the Sequencher software v.5.1. (Gene Codes Corporation).

2.6. Fluorescence in situ hybridization (FISH)

FISH analysis was performed using a dual-color fluorescence-labeled probe set for the *TRIB2* and *PRKCE* genes (Orange dUTP-labeled *TRIB2* and Green dUTP-labeled *PRKCE* probes) (Chromosome Science, Sapporo, Japan), as described previously [16]. Chromosomal localization and specificity of probe hybridization were validated on metaphase spreads of normal human lymphocytes.

2.7. Identification of driver genes and pathways

To identify cancer genes with mutations that drive the cancer phenotype and related pathways in the genome sequencing data, the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database was searched using the DrGap program (https://code.google.com/archive/p/drgap/) [17].

2.8. Real-time PCR

First-strand cDNA was synthesized from 1 μg of total RNA using a high-capacity cDNA reverse transcription kit (Life Technologies) in accordance with the manufacturer's instructions. Real-time PCR has performed as described previously [18]. Primers and probes sets were obtained from Applied Biosystems, and their Assay IDs are provided in Supplementary Table S2. The amplification reaction was performed in accordance with to the manufacturer's instructions (95 °C for 10 min followed by 40 cycles of 95 °C for 15 s, 50 °C for 2 min, and 60 °C for 1 min).

2.9. Statistical analysis

Overall survival was measured as the period from surgery to the date of death or last follow-up. All statistical analyses were performed using tools available in the R statistical package (version 3.6.1; http://www.r-project.org/). Differences at $P < 0.05$ were considered to be statistically significant.

2.10. Database deposition

Gene expression, WES, and RNA sequencing data have been deposited in the Gene Expression Omnibus (GEO) database with the accession numbers, GSE141755, GSE142190, and GSE142186, respectively. The deposits have been integrated into GSE142191.

3. Results

3.1. Clinicopathological characteristics of PC

The clinical and pathological characteristics of the 25 PC patients examined in this study are summarized in Table 1. There were 20 with TC (80%) and 5 with AC (20%). Nineteen cases (76%) were diagnosed at pathological stage **I** [according to the International Union Against Cancer (UICC) TNM Classification of Malignant Tumors, 8th edition (2017)], 5 cases (20 %) at stage II, and 2 (8%) cases at stage III. No patient had distant metastasis at the time of surgery or received systemic treatment after surgical resection. The follow-up periods ranged from 3 to 83 months (median follow-up, 49 months). Four (80 %) of the 5 patients with AC developed metastases, but one (5%) of the 20 patients with TC also developed metastases and one (20 %) of the 5 patients with AC (stage IB) did not developed metastasis within 5 years after surgical resection. This indicates that the histological classification of PC (TC or AC) is not sufficient for prediction of recurrence. Although no patients lacking histological lymphovascular tumor invasion developed recurrence, the presence of lymphovascular invasion was not predictive of postsurgical recurrence (Table 1).

3.2. Exon array-based expression profiling

The 25 PC samples were subjected to genome-wide transcriptome analysis using the GeneChip Human Exon 1.0 ST array. This exon array can detect mRNAs with low abundance as well as alternatively polyadenylated and spliced mRNA, because the probes are designed to hybridize to the entire sequences of transcripts [19]. In PCs that developed postsurgical metastasis, we identified 14 genes (*DENND1B, GRID1, CLMN, DENND1B, NRP1, SEL1L3, C5orf13, TNFRSF21, TES, STK39, MTHFD2, OPN3, MET,* and *HIST1H3C*) that were up-regulated and 71 that were down-regulated [> 2-fold, $p < 0.05$ (*t* test with no correction)] (Supplementary Tables S3 and S4). The 85 genes that were differentially expressed were clustered according to the similarity of their expression profiles (Fig. 1A), and the differential expression of representative genes was validated by real-time PCR in 5 cases that recurred (solid columns, Fig. 1B**)** and 5 cases that did not (clear columns, Fig. 1B). Patients with high expression of the *MET*, *TES*, and *STK39* genes showed a significantly unfavorable outcome (Fig. 1C).

3.3. Recurrent mutation of mucin genes

We performed the whole-exon sequencing of paired normal and tumor samples from 14 PC patients (10 TC and 4 AC). Under stringent selection criteria [14], we identified a total of 139 somatic alterations in 136 genes (Supplementary Table S5) [mean somatic mutation rate of 0.3 per megabase (Mb)]. The number of somatic alterations was associated with both Ki67 index [20] ($p < 0.01$, *t* test) and mitotic counts (per 2 mm^2) ($p < 0.01$, *t* test), but not with recurrence after surgery ($p = 0.187$, *t* test). Eighteen representative alterations were validated by dideoxynucleotide sequencing. Recurrent mutations were detected in 2 genes (*MUC6* and *SPTA1*) with a frequency of 21% (3/14) and 14% (2/

Table 1
Clinico-patholocial characteristics of 25 PC patients examined in this study.

		All patients ($n = 25$)		Recurrence (−) ($n = 20$)		Recurrence (+) ($n = 5$)		*p*-values*
Gender	Male	15	(60)	11	(44)	4	(16)	0.615
	Female	10	(40)	9	(36)	1	(4)	
Age	≥ 65	7	(28)	7	(28)	0	0	0.274
	< 65	18	(72)	13	(52)	5	(20)	
Hisologic subtype	Typical carcinoid	20	(80)	19	(76)	1	(4)	**0.0019**
	Atypical carcinoid	5	(20)	1	(4)	4	(16)	
Smoking status	Never smoked	17	(68)	14	(56)	3	(12)	1
	Current or former smoker	8	(32)	6	(24)	2	(8)	
Pathological stage	I	17	(68)	17	(68)	0	0	**0.0011**
	II and III	8	(32)	3	(12)	5	(20)	
Tumor size	≥3 cm	18	(72)	16	(64)	2	(8)	0.113
	< 3 cm	7	(28)	4	(16)	3	(12)	
Surgical margin (microscopic)	Negative	22	(88)	18	(72)	4	(16)	0.504
	Positive	3	(12)	2	(8)	1	(4)	
Lymphovascular invasion	Absent	16	(64)	16	(64)	0	0	**0.0024**
	Present	9	(36)	4	(16)	5	(20)	
Ki67 index	< 5%	21	(84)	18	(72)	3	(12)	0.166
	≥5%	4	(16)	2	(8)	2	(8)	
Mitosis	< 2	20	(80)	19	(76)	1	(4)	**0.0019**
	≥2	5	(20)	1	(4)	4	(16)	
Lymph node metastasis	Absent	19	(76)	17	(68)	2	(8)	0.0698
	Present	6	(24)	3	(12)	3	(12)	

p values of < 0.05 are shown in bold.
* Fisher's exact test.

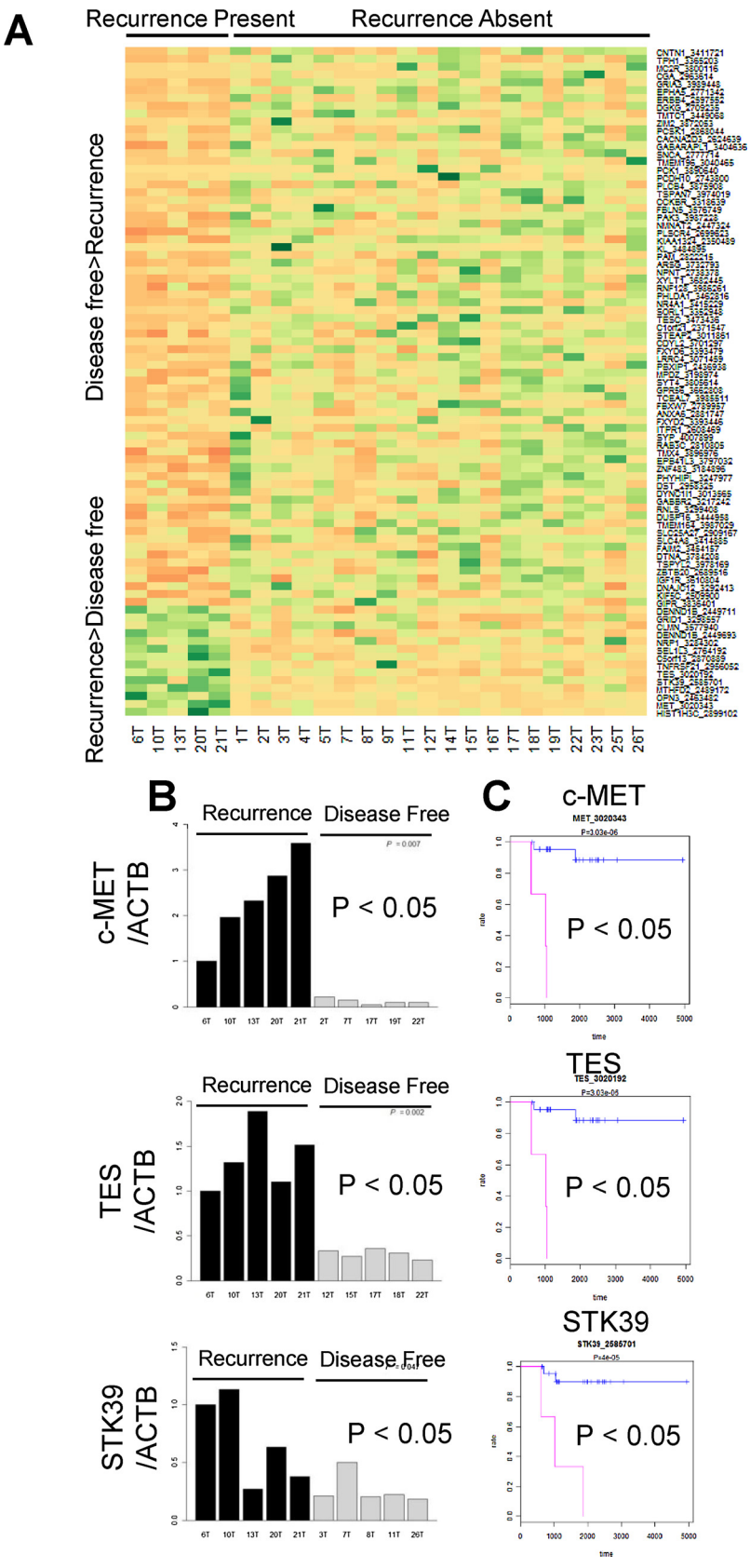


Fig. 1. Gene expression profiles associated with PC recurrence.
(A) Hierarchical clustering of 85 genes whose expression differed significantly ($p < 0.05$ and > 2-fold change) between PCs that developed metastasis ($n = 5$) and PCs that did not ($n = 20$).
(B) The expression levels of 3 representative genes (*MET*, *TES*, and *STK39*) significantly up-regulated in PC that developed metastasis were validated by real-time RT-PCR.
(C) Kaplan–Meier estimates of the overall survival of patients with PCs showing levels of *MET*, *TES*, *STK39* gene expression lower (blue) or higher (red) than the cut-off values with the lowest *p* values determined as described previously [41]. Differences between curves were evaluated using the log-rank test. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

14), respectively (Fig. 2). Interestingly, we also detected mutations in other mucin genes (*MUC2* and *MUC4*) in one sample each. In total, 5 (36%) of the 14 PCs were found to carry mutually exclusive *MUC* family gene mutations.

3.4. Pathway analysis of mutated genes

To identify the unique sets of genes associated with the pathogenesis of PC, the 136 mutated genes were evaluated using the DrGaP

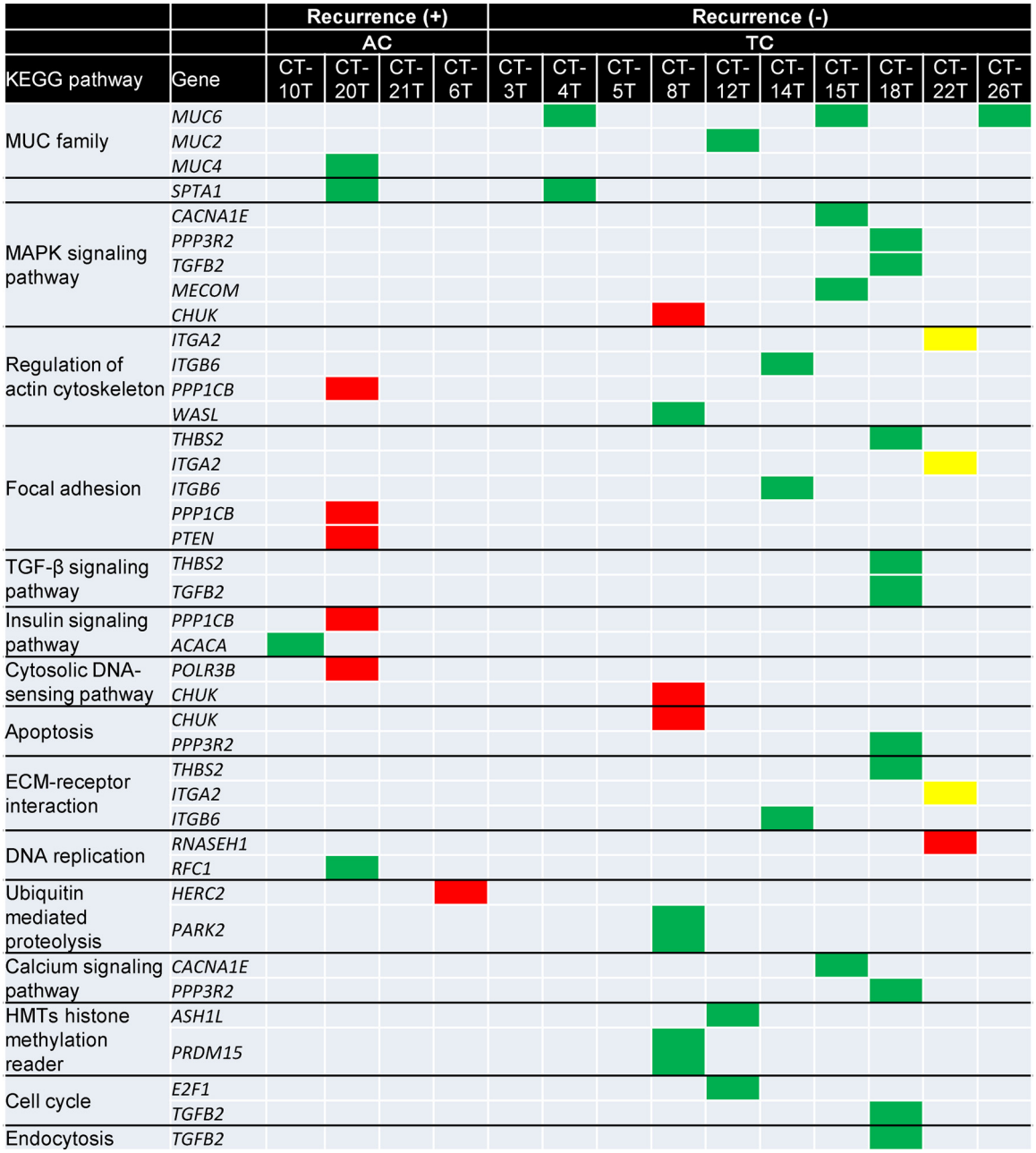


Fig. 2. Somatic mutations of PC.
One hundred thirty nine somatic mutations detected by WES were classified according to KEGG pathways. Green, non-synonymous mutations; Red, nonsense and frame-shift mutations; Yellow, splice site mutation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

program, which is designed to identify driver genes and pathways. Among the 136 genes, we found significant ($< 1.0 \times 10^{-5}$) enrichment of genes involved in MAPK signaling (*CACNA1E*, *PPP3R2*, *TGFB2*, *MECOM*, and *CHUK*), regulation of the actin cytoskeleton (*ITGA2, ITGB6, PPP1CB*, and *WASL*), focal adhesion (*LAMC1, THBS2, ITGA2, ITGB6, PPP1CB*, and *PTEN*), TGF-β signaling (*THBS2* and *TGFB2*), cytosolic DNA-sensing, and insulin signaling (Table 2). There were significant differences in mutation type (non-synonymous, nonsense and frame-shift, or splice site mutation) and frequency among the cases, but none of the somatic mutations (Supplementary Table S5) or pathways (Fig. 2) was found to be associated with PC recurrence.

3.5. Chromosomal rearrangements in PC

We next performed RNA sequencing of tumor samples from 6 PC patients (including 4 who developed postsurgical metastasis) and identified a total of 8 novel fusion transcripts (Supplementary Table S6). All of the 8 fusion genes were subsequently validated by RT-PCR across their relevant exon-exon boundaries and by dideoxynucleotide sequencing. The inter- and intrachromosomal translocations responsible frequently involved chromosomes 1, 2, 9, 12, 17, and 20 (Fig. 3A). Only the *SUZ12* gene in chromosome 7 was recurrently involved in the sites of chromosomal translocation (Supplementary Table S6). The detection of fusion genes was associated with mitotic counts ($p = 0.032$, t test), but not with Ki67 index ($p = 0.14$, t test) (Supplementary Table S7). It is noteworthy, however, that all of the 8 fusion

Table 2
Pathway analysis of genes mutated in PC.

KEGG pathway	KEGG ID	*p*-values	Number of mutated genes	Genes
MAPK signaling pathway	hsa04010	5.11E − 08	6	*CACNA1B, CACNA1E, PPP3R2, TGFB2, MECOM, CHUK*
Regulation of actin cytoskeleton	hsa04810	1.01E − 07	4	*ITGA2, ITGB6, PPP1CB, WASL*
Focal adhesion	hsa04510	1.13E − 07	6	*LAMC1, THBS2, ITGA2, ITGB6, PPP1CB, PTEN*
TGF-β signaling pathway	hsa04350	1.33E − 06	2	*THBS2, TGFB2*
Insulin signaling pathway	hsa04910	3.94E − 06	2	*PPP1CB, ACACA*
Cytosolic DNA-sensing pathway	hsa04623	4.79E − 06	2	*POLR3B, CHUK*
Apoptosis	hsa04210	1.15E − 05	2	*CHUK, PPP3R2*
ECM-receptor interaction	hsa04512	1.55E − 05	4	*LAMC1, THBS2, ITGA2, ITGB6*
DNA replication	hsa03030	1.70E − 05	2	*RNASEH1, RFC1*
Ubiquitin mediated proteolysis	hsa04120	0.00157	2	*HERC2, PARK2*
Calcium signaling pathway	hsa04020	1.58E − 03	3	*CACNA1B, CACNA1E, PPP3R2*
HMT histone methylation reader	u0002	0.00254	2	*ASH1L, PRDM15*
Cell cycle	hsa04110	0.00727	2	*E2F1, TGFB2*
Endocytosis	hsa04144	0.01207	2	*TGFB2, ARAP2*

Abbreviations: MAPK, mitogen-activated protein kinase; TGF-β, transforming growth factor-β; ECM, extracellular matrix; HMT, histone methyltransferase.

genes were detected only in 4 cases that had developed postsurgical metastasis (Supplementary Table S7).

Of the 8 fusion transcripts, only *TRIB2-PRKCE* involved a tyrosine kinase gene. Protein kinase C epsilon (*PRKCE*), a member of the novel protein kinase C (PKC) family, plays key roles in the mitogenesis and survival of normal and cancer cells [21]. The *TRIB2-PRKCE* transcript was an in-frame fusion between exon 1 of *TRIB2* and exon 2 of *PRKCE* (Fig. 3B) and deduced to encode a truncated protein lacking the N-terminal C2 regulatory domain of *PRKCE* (Fig. 3C). The *TRIB2* and *PRKCE* genes are located on both chromosomes 2 (Fig. 3D). The intrachromosomal rearrangement involving the *TRIB2* and *PRKCE* loci was confirmed by FISH (Supplementary Fig. S1A). The *TRIB2* gene has 15 exons (exons 1–15), but the expression of exon 1 was suppressed (Supplementary Fig. S1B) probably by the chromosomal translocation.

4. Discussion

PC is a rare, low- or intermediate- grade tumor that contains significantly fewer genetic abnormalities in comparison with other malignancies [2,22], and for this reason no definite therapeutic target molecule has so far been identified. Fernandez-Cuesta et al. reported a mean somatic mutation rate of 0.4 (comparable to this study) with frequent mutations in chromatin remodeling genes, such as *MEN1* and *ARID1A*. They also identified a case with chromothripsis. The case showed intensely clustered genomic structural alterations [copy number variations (CNV) and chromosomal rearrangements] in chromosomes 3, 12 and 13, but chromosomal translocation was infrequent in PC and only a few fusion transcripts were detectable in other cases. In this study, we were able to detect chromosomal rearrangements in all 4 cases that had developed postsurgical metastasis (Supplementary Table S7). Chromosomal instability (CIN) has been considered a hallmark of cancers with unfavorable outcome. Detection of fusion genes may represent the CIN status of PC and indicate a high risk of recurrence.

Among the fusion genes detected in this study, only *TRIB2-PRKCE* was deduced to encode an in-frame fusion protein (Fig. 3 and Supplementary Table S6). However, fusion transcripts involving the *PRKCE*, *SUZ12*, and *SFPQ* genes have been repeatedly reported in other malignancies (Supplementary Table S8), indicating their involvement in carcinogenesis. PKC is a family of serine/threonine specific protein kinases that can be activated by calcium and the second messenger diacylglycerol. PKC family members phosphorylate a wide variety of protein targets and are known to be involved in diverse cellular signaling pathways [23]. Genomic alterations in PKC family members have been identified in several cancers [21,23–27]. Fusion transcripts involving the *PRKCE* gene have been reported in colorectal, lung, and breast cancers (Supplementary Table S8) [9,25]. The *JAZF1/SUZ12* (suppressor of Zeste 12) gene translocation is the most frequent chromosomal abbreviation in endometrial stromal sarcoma (ESS) and considered to be its cytogenetic hallmark (Supplementary Table S8) [28,29]. *SUZ12* is required for the gene expression program of embryonic stem (ES) cell differentiation [30]. Expression of EZH2, which is complexed with at least two of its non-catalytic partners, EED and SUZ12, is required in order for Ewing tumors to maintain their cancer stemness [31]. *SFPQ* (splicing factor, proline- and glutamine-rich)/*PSF* (PTB-associated splicing factor) is a common fusion partner of *TFE3* in papillary renal cell carcinoma [32].

We detected recurrent mutations in the mucin genes (including *MUC2*, *MUC4*, and *MUC6*) and the *SPTA1* gene. The *MUC4* mutation was detected in exon 2. It has been reported that *MUC4* mutations are concentrated in the exon in lung cancer [33]. The high frequency of mutations appeared to indicate involvement in pathogenesis. *MUC4* encodes a membrane-spanning mucin glycoprotein on the ciliated airway mucosal surface [34,35] and plays an important role in the proliferation and differentiation of epithelial cells by regulating the phosphorylation of ErbB2 and expression of the cyclin-dependent kinase inhibitor p27 [36–38]. *MUC6* encoding gastric mucin with a mucoprotective function was mutated in 9.6 % of microsatellite-stable (MSS) and 18.2 % of microsatellite-unstable (MSI) gastric cancers [39]. Down-regulation of MUC6 protein was reported to correlate with advanced tumor stage and poor prognosis in gastric cancer [40]. Functional characterization of *MUC* gene mutations may provide an avenue by which they can be exploited as potential therapeutic targets. Genetic alterations affecting the *SPTA1* gene were reported to be present in 11.5 % of small cell lung cancers [41], but its pathological significance has not been established.

The present and previous studies showed the different spectra of somatic mutations. Table 3 lists preceding studies in which NGS was employed. We speculate that the difference may have been due to the small number of patients we included, differences in patient ethnicity, and/or differences in the sequencing and bioinformatical methods employed.

In this study we identified novel somatic mutations and chromosomal rearrangements by examining clinically aggressive PC cases that had developed postsurgical metastasis. Unfortunately, none of the genetic alterations except for the *TRIB2-PRKCE* fusion transcript was considered technically actionable. Further investigation through a different molecular approach would be necessary to uncover targetable genetic alterations in such patients. The functional relevance of 14 genes up-regulated in PCs that developed recurrence (Supplementary Table S3) is now under investigation.

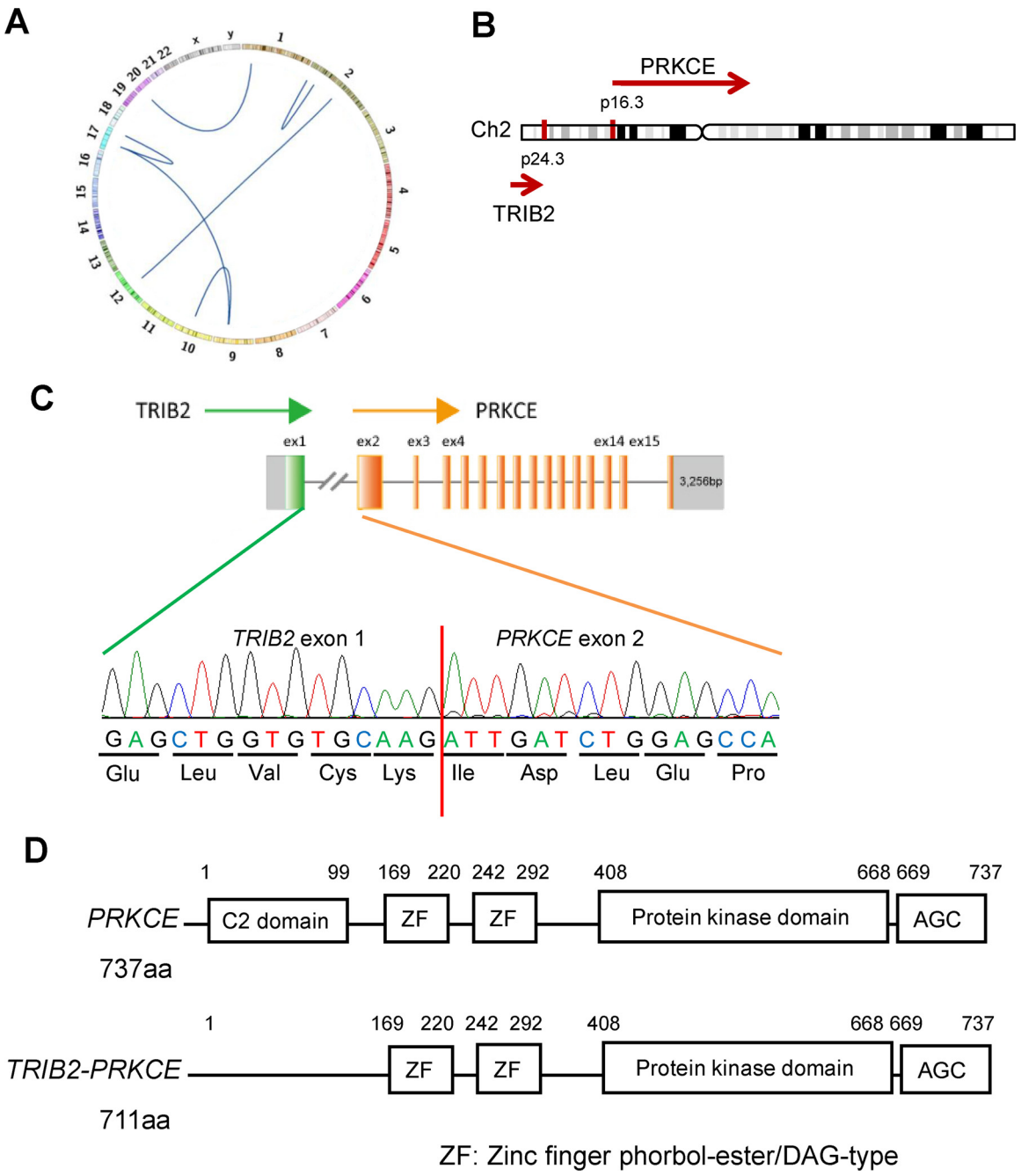


Fig. 3. Identification of the *TRIB2-PRKCE* fusion gene.
(A) Intra/interchromosomal translocation detected in PC.
(B) Ideogram showing the intrachromosomal translocation of *TRIB2* (chromosomes 2 at p24.3) and *PRKCE* (chromosome 2 at p16.3).
(C) Nucleotide and deduced amino acid sequences at the break/fusion point of the *TRIB2-PRKCE* transcript.
(D) Schematic representation of the deduced domain structure of TRIB2-PRKCE (TOP) and wild-type PRKCE (BOTTOM) proteins. Wild-type PRKCE contains the C2 [amino acids 1–99], zinc finger phorbol-ester/DAG-type [169–292], protein kinase [408–668], and AGC-kinase C-terminal [669–737] domains.

Author Statement

Akihiko Miyanaga: Conceptualization, Methodology, Investigation, Data Curation, Visualization, Writing - original draft.
Mari Masuda: Conceptualization, Methodology, Investigation, Data Curation, Writing - Review & Editing.
Noriko Motoi: Investigation, Resources.
Koji Tsuta: Investigation, Resources.
Yuka Nakamura: Administrative, technical, or material support.
Nobuhiko Nishijima: Data Curation.
Shun-ichi Watanabe: Resources, Data Curation.
Hisao Asamura: Resources, Data Curation.
Akihiko Tsuchida: Writing - Review & Editing.
Masahiro Seike: Writing - Review & Editing. Akihiko Gemma: Writing - Review & Editing.
Tesshi Yamada: Conceptualization, Methodology, Investigation, Data Curation, Visualization, Writing - Review & Editing, Supervision.

Declaration of Competing Interest

The authors have no potential conflicts of interest to disclose.

Table 3
Preceding NGS studies of PC.

Reference	Tumor materials	No. of patients	Country	Sequencing method/NGS platform	Number of mutations	Recurrent mutations	Additional analyses
Fernandez-Cuesta et al.*	Fresh frozen samples	44	USA	WES/HiSeq 2000	529 non-synonymous mutations in 494 genes	Chromatin remodeling genes (including *MEN1*, *ARID1A*, *EIF1AX*)	RNA-seq (n = 39), SNP array (n = 29)
Vollbrecht et al.**	FFPE samples	17 TC,17 AC, 19 LCNEC, 17 SCLC	Germany	TruSeq Amplicon-Cancer panel/MiSeq	48 genes	None	None
Boora et al.***	4 cell lines	None	–	WES/HiSeq 2000	1425 non-synonymous SNV, 260 SNV Mutation rate of 8 per Mb	*BRCA1*, *PTCH1*, *STK11*, *APC2*	None
Asiedu et al.****	3 cell lines established from patients	3 TC	USA	WES/HiSeq 2000	14 genes (*ACVR2A*, *MUC2*, *BCLAF1*, *SEMA3B*, *RETNLB*, *MUC4*, *NOP16*, *PODXL*, *DCP1B*, *EP400*, *HRC*)	None	RNA-seq
Asiedu et al.*****	Fresh frozen samples	14 TC, 6 AC	USA	HiSeq 2500	126 cancer driver mutations	*ATP1A2, CNNM1, MACF1, RAB38, NF1, RAD51C, TAF1L, EPHB2*, *POLR3B*, *AGFG1*	Genotyping
Alcala et al.******	Fresh frozen samples	36 TC, 27 AC	USA	WES/HiSeq 2000	485 mutations	Chromatin remodeling genes (including *MEN1*, *ARID1A*, *EIF1AX*), *ATM*, *PSIP1* and *ROBO1*	RNA-seq, Targeted sequencing
Laddha et al.*******	Fresh frozen samples	17 TC, 13 AC	USA	Targeted cancer gene panel DNA sequencing	37 mutations	histone modification/chromatin remodeling genes (*MEN1, ARID1A, KMT2A, KMT2C, KMT2D,* and *SMARCA4*) and DNA repair pathways	RNA-seq, DNA methylation array
Simbolo et al.********	FFPE samples	35 AC	Italy	Ion Torrent platform	67 mutations	*TP53, RB1, MEN1*	None
Current study	Fresh frozen samples	10 TC, 4 AC	Japan	WES/HiSeq 2000	139 mutations in 136 genes	*MUC6*, *SPTA1*	RNA-seq (n = 6), Exon array (n = 25)

Abbreviations: FFPE, formalin fixed paraffin embedded; NGS, next-generation sequencing; TC, typical carcinoids; AC, atypical carcinoids; WES, whole-exome sequencing; RNA-seq, RNA sequencing; SNP, single nucleotide polymorphism; LCNEC, large cell neuroendocrine; carcinoma; SCLC, small cell lung carcinoma.

* Nat Commun. 2014 Mar 27;5:3518.
** Br J Cancer. 2015 Dec 22;113(12):1704 – 11.
*** Cancer Genet. 2015 Jul-Aug;208(7–8):374 – 381.
**** J Thorac Oncol. 2014 Dec;9(12):1763 – 71.
***** Clin Cancer Res. 2018 Apr 1;24(7):1691 – 1704.
****** Nat Commun. 2019 Aug 20;10(1):3407.
******* Clin Cancer Res. 2019 Sep 1;79(17):4339 – 4347.
******** J Thorac Oncol. 2019 Sep;14(9):1651 – 1661.

Acknowledgments

This study was supported by the National Cancer Center Research and Development Fund (26-A-13 and26-A-5to T. Yamada and 30-A-2 to M. Masuda), the Acceleration Transformative Research for Medical Innovation (ACT-MS) program of the Japan Agency for Medical Research and Development (AMED) (16im0210804h0001 to T. Yamada), the Kobayashi Foundation for Cancer Research (to T. Yamada), a KAKENHI Grant-in-Aid for Scientific Research (26461168 to A. Miyanaga), a KAKENHI Grant-in-Aid for Challenging Exploratory Research (19H05566 to T. Yamada and 16K14627 to M. Masuda), a Grant-in Aid for Scientific Research (B) (17H03603 to M. Masuda) from the Japan Society for the Promotion of Science (JSPS), a cancer research grant from the Foundation for Promotion of Cancer Research in Japan (to M. Masuda), the Project Mirai Cancer Research Grant from the Japan Cancer Society(to M. Masuda), and a research grant from the Princess Takamatsu Cancer Research Fund (to M. Masuda). The authors would like to thank K. Hayashi M.S. for technical assistance and the Fundamental Innovative Oncology Core (FIOC) of the National Cancer Center Research Institute (Tokyo, Japan) for sequence data analysis.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.lungcan.2020.03.027.

References

[1] W.D. Travis, E. Brambilla, A.G. Nicholson, Y. Yatabe, J.H. Austin, M.B. Beasley, L.R. Chirieac, S. Dacic, E. Duhig, D.B. Flieder, K. Geisinger, F.R. Hirsch, Y. Ishikawa, K.M. Kerr, M. Noguchi, G. Pelosi, C.A. Powell, M.S. Tsao, I. Wistuba, The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification, J. Thorac. Oncol. 10 (9) (2015) 1243–1260.

[2] L. Fernandez-Cuesta, M. Peifer, X. Lu, R. Sun, L. Ozretic, D. Seidel, T. Zander, F. Leenders, J. George, C. Muller, I. Dahmen, B. Pinther, G. Bosco, K. Konrad, J. Altmuller, P. Nurnberg, V. Achter, U. Lang, P.M. Schneider, M. Bogus, A. Soltermann, O.T. Brustugun, A. Helland, S. Solberg, M. Lund-Iversen, S. Ansen, E. Stoelben, G.M. Wright, P. Russell, Z. Wainer, B. Solomon, J.K. Field, R. Hyde, M.P. Davies, L.C. Heukamp, I. Petersen, S. Perner, C.M. Lovly, F. Cappuzzo, W.D. Travis, J. Wolf, M. Vingron, E. Brambilla, S.A. Haas, R. Buettner, R.K. Thomas, Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids, Nat. Commun. 5 (2014) 3518.

[3] J.C. Yao, M. Hassan, A. Phan, C. Dagohoy, C. Leary, J.E. Mares, E.K. Abdalla, J.B. Fleming, J.N. Vauthey, A. Rashid, D.B. Evans, One hundred years after "carcinoid": epidemiology of and prognostic factors for neuroendocrine tumors in 35,825 cases in the United States, J. Clin. Oncol. 26 (18) (2008) 3063–3072.

[4] H. Asamura, T. Kameya, Y. Matsuno, M. Noguchi, H. Tada, Y. Ishikawa, T. Yokose, S.X. Jiang, T. Inoue, K. Nakagawa, K. Tajima, K. Nagai, Neuroendocrine neoplasms of the lung: a prognostic spectrum, J. Clin. Oncol. 24 (1) (2006) 70–76.

[5] L.J. Wirth, M.R. Carter, P.A. Janne, B.E. Johnson, Outcome of patients with pulmonary carcinoid tumors receiving chemotherapy or chemoradiotherapy, Lung Cancer 44 (2) (2004) 213–220.

[6] M.H. Kulke, H. Kim, K. Stuart, J.W. Clark, D.P. Ryan, M. Vincitore, R.J. Mayer, C.S. Fuchs, A phase II study of docetaxel in patients with metastatic carcinoid tumors, Cancer Invest. 22 (3) (2004) 353–359.

[7] A. Lal, H. Chen, Treatment of advanced carcinoid tumors, Curr. Opin. Oncol. 18 (1) (2006) 9–15.

[8] E.M. Bertino, P.D. Confer, J.E. Colonna, P. Ross, G.A. Otterson, Pulmonary neuroendocrine/carcinoid tumors: a review article, Cancer 115 (19) (2009) 4434–4441.

[9] N. Stransky, E. Cerami, S. Schalm, J.L. Kim, C. Lengauer, The landscape of kinase fusions in cancer, Nat. Commun. 5 (2014) 4846.

[10] M. Peifer, L. Fernandez-Cuesta, M.L. Sos, J. George, D. Seidel, L.H. Kasper, et al., Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer, Nat. Genet. 44 (10) (2012) 1104–1110.

[11] M. Imielinski, A.H. Berger, P.S. Hammerman, B. Hernandez, T.J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M.S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansen, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B.E. Johnson, P.A. Janne, V.A. Miller, W. Pao, W.D. Travis, H.I. Pass, S.B. Gabriel, E.S. Lander, R.K. Thomas, L.A. Garraway, G. Getz, M. Meyerson, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing, Cell 150 (6) (2012) 1107–1120.

[12] A genomics-based classification of human lung tumors, Sci. Transl. Med. 5 (209) (2013) 209ra153.

[13] G. Lou, X. Yu, Z. Song, Molecular profiling and survival of completely resected primary pulmonary neuroendocrine carcinoma, Clin. Lung Cancer 18 (3) (2017) e197–e201.

[14] A. Miyanaga, M. Masuda, K. Tsuta, K. Kawasaki, Y. Nakamura, T. Sakuma, H. Asamura, A. Gemma, T. Yamada, Hippo pathway gene mutations in malignant mesothelioma: revealed by RNA and targeted exon sequencing, J. Thorac. Oncol. 10 (5) (2015) 844–851.

[15] A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M.G. Sun, M. Griffith, A. Heravi Moussavi, J. Senz, N. Melnyk, M. Pacheco, M.A. Marra, M. Hirst, T.O. Nielsen, S.C. Sahinalp, D. Huntsman, S.P. Shah, deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data, PLoS Comput. Biol. 7 (5) (2011) e1001138.

[16] R. Noro, K. Honda, K. Tsuta, G. Ishii, A.M. Maeshima, N. Miura, K. Furuta, T. Shibata, H. Tsuda, A. Ochiai, T. Sakuma, N. Nishijima, A. Gemma, H. Asamura, K. Nagai, T. Yamada, Distinct outcome of stage I lung adenocarcinoma with ACTN4 cell motility gene amplification, Ann. Oncol. 24 (10) (2013) 2594–2600.

[17] X. Hua, H. Xu, Y. Yang, J. Zhu, P. Liu, Y. Lu, DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies, Am. J. Hum. Genet. 93 (3) (2013) 439–451.

[18] R. Satow, M. Shitashige, Y. Kanai, F. Takeshita, H. Ojima, T. Jigami, K. Honda, T. Kosuge, T. Ochiya, S. Hirohashi, T. Yamada, Combined functional genome survey of therapeutic targets for hepatocellular carcinoma, Clin. Cancer Res. 16 (9) (2010) 2518–2528.

[19] P.J. Gardina, T.A. Clark, B. Shimada, M.K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, Y. Turpaz, Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array, BMC Genomics 7 (2006) 325.

[20] J.K.T. Dermawan, C.F. Farver, The role of histologic grading and Ki-67 index in predicting outcomes in pulmonary carcinoid tumors, Am. J. Surg. Pathol. 44 (2) (2020) 224–231.

[21] E.M. Griner, M.G. Kazanietz, Protein kinase C and other diacylglycerol effectors in cancer, Nat. Rev. Cancer 7 (4) (2007) 281–294.

[22] C. Vollbrecht, R. Werner, R.F. Walter, D.C. Christoph, L.C. Heukamp, M. Peifer, B. Hirsch, L. Burbat, T. Mairinger, K.W. Schmid, J. Wohlschlaeger, F.D. Mairinger, Mutational analysis of pulmonary tumours with neuroendocrine features using targeted massive parallel sequencing: a comparison of a neglected tumour group, Br. J. Cancer 113 (12) (2015) 1704–1711.

[23] M.A. Gorin, Q. Pan, Protein kinase C epsilon: an oncogene and emerging tumor biomarker, Mol. Cancer 8 (2009) 9.

[24] M.C. Caino, C. Lopez-Haber, J. Kim, D. Mochly-Rosen, M.G. Kazanietz, Proteins kinase Cvarepsilon is required for non-small cell lung carcinoma growth and regulates the expression of apoptotic genes, Oncogene 31 (20) (2012) 2593–2600.

[25] P.J. Stephens, D.J. McBride, M.L. Lin, I. Varela, E.D. Pleasance, J.T. Simpson, L.A. Stebbings, C. Leroy, S. Edkins, L.J. Mudie, C.D. Greenman, M. Jia, C. Latimer, J.W. Teague, K.W. Lau, J. Burton, M.A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A.M. Sieuwerts, J.W. Martens, D.P. Silver, A. Langerod, H.E. Russnes, J.A. Foekens, J.S. Reis-Filho, L. van' t Veer, A.L. Richardson, A.L. Borresen-Dale, P.J. Campbell, P.A. Futreal, M.R. Stratton, Complex landscapes of somatic rearrangement in human breast cancer genomes, Nature 462 (7276) (2009) 1005–1010.

[26] J.S. Seo, Y.S. Ju, W.C. Lee, J.Y. Shin, J.K. Lee, T. Bleazard, J. Lee, Y.J. Jung, J.O. Kim, J.Y. Shin, S.B. Yu, J. Kim, E.R. Lee, C.H. Kang, I.K. Park, H. Rhee, S.H. Lee, J.I. Kim, J.H. Kang, Y.T. Kim, The transcriptional landscape and mutational profile of lung adenocarcinoma, Genome Res. 22 (11) (2012) 2109–2119.

[27] P. Kim, X. Zhou, FusionGDB: fusion gene annotation DataBase, Nucleic Acids Res. 47 (D1) (2019) D994–D1004.

[28] A. Hrzenjak, JAZF1/SUZ12 gene fusion in endometrial stromal sarcomas, Orphanet J. Rare Dis. 11 (2016) 15.

[29] J.I. Koontz, A.L. Soreng, M. Nucci, F.C. Kuo, P. Pauwels, H. van Den Berghe, P. Dal Cin, J.A. Fletcher, J. Sklar, Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors, Proc. Natl. Acad. Sci. U.S.A. 98 (11) (2001) 6348–6353.

[30] D. Pasini, A.P. Bracken, J.B. Hansen, M. Capillo, K. Helin, The polycomb group protein Suz12 is required for embryonic stem cell differentiation, Mol. Cell. Biol. 27 (10) (2007) 3769–3779.

[31] S. Burdach, S. Plehm, R. Unland, U. Dirksen, A. Borkhardt, M.S. Staege, C. Muller-Tidow, G.H. Richter, Epigenetic maintenance of stemness and malignancy in peripheral neuroectodermal tumors by EZH2, Cell Cycle 8 (13) (2009) 1991–1996.

[32] M. Mathur, S. Das, H.H. Samuels, PSF-TFE3 oncoprotein in papillary renal cell carcinoma inactivates TFE3 and p53 through cytoplasmic sequestration, Oncogene 22 (32) (2003) 5031–5044.

[33] S. Kumar, E. Cruz, S. Joshi, A. Patel, R. Jahan, S.K. Batra, M. Jain, Genetic variants of mucins: unexplored conundrum, Carcinogenesis 38 (7) (2017) 671–679.

[34] M. Kesimer, C. Ehre, K.A. Burns, C.W. Davis, J.K. Sheehan, R.J. Pickles, Molecular organization of the mucins and glycocalyx underlying mucus transport over mucosal surfaces of the airways, Mucosal Immunol. 6 (2) (2013) 379–392.

[35] M. Ali, E.P. Lillehoj, Y. Park, Y. Kyo, K.C. Kim, Analysis of the proteome of human airway epithelial secretions, Proteome Sci. 9 (2011) 4.

[36] S. Jepson, M. Komatsu, B. Haq, M.E. Arango, D. Huang, C.A. Carraway, K.L. Carraway, Muc4/sialomucin complex, the intramembrane ErbB2 ligand, induces specific phosphorylation of ErbB2 and enhances expression of p27(kip), but does not activate mitogen-activated kinase or protein kinaseB/Akt pathways, Oncogene 21 (49) (2002) 7524–7532.

[37] N. Jonckheere, M. Perrais, C. Mariette, S.K. Batra, J.P. Aubert, P. Pigny, I. Van Seuningen, A role for human MUC4 mucin gene, the ErbB2 ligand, as a target of TGF-β in pancreatic carcinogenesis, Oncogene 23 (34) (2004) 5729–5738.

[38] A.P. Singh, P. Chaturvedi, S.K. Batra, Emerging roles of MUC4 in cancer: a novel

target for diagnosis and therapy, Cancer Res. 67 (2) (2007) 433–436.
[39] K. Wang, S.T. Yuen, J. Xu, S.P. Lee, H.H. Yan, S.T. Shi, H.C. Siu, S. Deng, K.M. Chu, S. Law, K.H. Chan, A.S. Chan, W.Y. Tsui, S.L. Ho, A.K. Chan, J.L. Man, V. Foglizzo, M.K. Ng, A.S. Chan, Y.P. Ching, G.H. Cheng, T. Xie, J. Fernandez, V.S. Li, H. Clevers, P.A. Rejto, M. Mao, S.Y. Leung, Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer, Nat. Genet. 46 (6) (2014) 573–582.
[40] H. Zheng, H. Takahashi, T. Nakajima, Y. Murai, Z. Cui, K. Nomoto, K. Tsuneyama, Y. Takano, MUC6 down-regulation correlates with gastric carcinoma progression and a poor prognosis: an immunohistochemical study with tissue microarrays, J. Cancer Res. Clin. Oncol. 132 (12) (2006) 817–823.
[41] J. Hu, Y. Wang, Y. Zhang, Y. Yu, H. Chen, K. Liu, M. Yao, K. Wang, W. Gu, T. Shou, Comprehensive genomic profiling of small cell lung cancer in Chinese patients and the implications for therapeutic potential, Cancer Med. 8 (9) (2019) 4338–4347.

Method

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet[1,†], Britta Velten[2,†], Damien Arnol[1], Sascha Dietrich[3], Thorsten Zenz[3,4,5], John C Marioni[1,6,7], Florian Buettner[1,8,*], Wolfgang Huber[2,**] & Oliver Stegle[1,2,***]

Abstract

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and *ex vivo* drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

Keywords data integration; dimensionality reduction; multi-omics; personalized medicine; single-cell omics

Subject Categories Computational Biology; Genome-Scale & Integrative Biology; Methods & Resources

DOI 10.15252/msb.20178124 | Received 27 November 2017 | Revised 28 May 2018 | Accepted 29 May 2018

Mol Syst Biol. (2018) 14: e8124

Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenome profiling (Hasin *et al*, 2017). Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omics profiling is increasingly applied across biological domains, including cancer biology (Gerstung *et al*, 2015; Iorio *et al*, 2016; Mertins *et al*, 2016; Cancer Genome Atlas Research Network, 2017), regulatory genomics (Chen *et al*, 2016), microbiology (Kim *et al*, 2016) or host-pathogen biology (Soderholm *et al*, 2016). Most recent technological advances have also enabled performing multi-omics analyses at the single-cell level (Macaulay *et al*, 2015; Angermueller *et al*, 2016; Guo *et al*, 2017; Clark *et al*, 2018; Colomé-Tatché & Theis, 2018). A common aim of such applications is to characterize heterogeneity between samples, as manifested in one or several of the data modalities (Ritchie *et al*, 2015). Multi-omics profiling is particularly appealing if the relevant axes of variation are not known *a priori*, and hence may be missed by studies that consider a single data modality or targeted approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus mapping, where large numbers of association tests are performed between individual genetic variants and gene expression levels (GTEx Consortium, 2015) or epigenetic marks (Chen *et al*, 2016). While eminently useful for variant annotation, such association studies are inherently *local* and do not provide a coherent global map of the molecular differences between samples. A second strategy is the use of kernel- or graph-based methods to combine different

1 European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK
2 European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
3 Heidelberg University Hospital, Heidelberg, Germany
4 German Cancer Research Center (dkfz) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
5 Germany & Hematology, University Hospital Zurich and University of Zurich, Zurich, Switzerland
6 Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
7 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
8 Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany
*Corresponding author. Tel: +49 89 23742560; E-mail: fbuettner.phys@gmail.com
**Corresponding author. Tel: +49 6221 387 8823; E-mail: wolfgang.huber@embl.de
***Corresponding author. Tel: +49 6221 3878190; E-mail: oliver.stegle@embl.de
†These authors contributed equally to this work

data types into a common similarity network between samples (Lanckriet *et al*, 2004; Wang *et al*, 2014); however, it is difficult to pinpoint the molecular determinants of the resulting graph structure. Related to this, there exist generalizations of other clustering methods to reconstruct discrete groups of samples based on multiple data modalities (Shen *et al*, 2009; Mo *et al*, 2013).

A key challenge that is not sufficiently addressed by these approaches is interpretability. In particular, it would be desirable to reconstruct the underlying factors that drive the observed variation across samples. These could be continuous gradients, discrete clusters or combinations thereof. Such factors would help in establishing or explaining associations with external data such as phenotypes or clinical covariates. Although factor models that aim to address this have previously been proposed (e.g. Meng *et al*, 2014, 2016; Tenenhaus *et al*, 2014; preprint: Singh *et al*, 2018), these methods either lack sparsity, which can reduce interpretability, or require a substantial number of parameters to be determined using computationally demanding cross-validation or post hoc. Further challenges faced by existing methods are computational scalability to larger data sets, handling of missing values and non-Gaussian data modalities, such as binary readouts or count-based traits.

Results

We present Multi-Omics Factor Analysis (MOFA), a statistical method for integrating multiple modalities of omics data in an unsupervised fashion. Intuitively, MOFA can be viewed as a versatile and statistically rigorous generalization of principal component analysis (PCA) to multi-omics data. Given several data matrices with measurements of multiple omics data types on the same or on partially overlapping sets of samples, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors (Fig 1A). These learnt factors capture major sources of variation across data modalities, thus facilitating the identification of continuous molecular gradients or discrete subgroups of samples. The inferred factor loadings can be sparse, thereby facilitating the linkage between the factors and the most relevant molecular features. Importantly, MOFA disentangles to what extent each factor is unique to a single data modality or is manifested in multiple modalities (Fig 1B), thereby revealing shared axes of variation between the different omics layers. Once trained, the model output can be used for a range of downstream analyses, including visualization, clustering and classification of samples in the low-dimensional space(s) spanned by the factors, as well as the automated annotation of factors using (gene set) enrichment analysis, the identification of outlier samples and the imputation of missing values (Fig 1B).

Technically, MOFA builds upon the statistical framework of group Factor Analysis (Virtanen *et al*, 2012; Khan *et al*, 2014; Klami *et al*, 2015; Bunte *et al*, 2016; Zhao *et al*, 2016; Leppäaho & Kaski, 2017), which we have adapted to the requirements of multi-omics studies (Materials and Methods): (i) fast inference based on a variational approximation, (ii) inference of sparse solutions facilitating interpretation, (iii) efficient handling of missing values and (iv) flexible combination of different likelihood models for each data modality, which enables integrating diverse data types such as binary-, count- and continuous-valued data. The relationship of MOFA to previous approaches (Shen *et al,* 2009; Virtanen *et al,* 2012; Mo *et al,* 2013; Klami *et al,* 2015; Remes *et al,* 2015; Bunte *et al,* 2016; Hore *et al,* 2016; Zhao *et al,* 2016; Leppáaho & Kaski, 2017) is discussed in Materials and Methods and Appendix Table S3.

MOFA is implemented as well-documented open-source software and comes with tutorials and example workflows for different application domains (Materials and Methods). Taken together, these functionalities provide a powerful and versatile tool for disentangling sources of variation in multi-omics studies.

Model validation and comparison on simulated data

First, to validate MOFA, we simulated data from its generative model, varying the number of views, the likelihood models, the number of latent factors and other parameters (Materials and Methods, Appendix Table S1). We found that MOFA was able to accurately reconstruct the latent dimension, except in settings with large numbers of factors or high proportions of missing values (Appendix Fig S1). We also found that models that account for non-Gaussian observations improved the fit when simulating binary or count data (Appendix Figs S2 and S3).

We also compared MOFA to two previously reported latent variable models for multi-omics integration: GFA (Leppäaho & Kaski, 2017) and iCluster (Mo *et al*, 2013). Over a range of simulations, we observed that GFA and iCluster tended to infer redundant factors (Appendix Fig S4) and were less accurate in recovering patterns of shared factor activity across views (Appendix Fig S5). MOFA is also computationally more efficient than these existing methods (Fig EV1). For example, the training on the CLL data, which we consider next, required 25 min using MOFA versus 34 h with GFA and 5–6 days with iCluster.

Application to chronic lymphocytic leukaemia

We applied MOFA to a study of chronic lymphocytic leukaemia (CLL), which combined *ex vivo* drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays (Dietrich *et al*, 2018; Fig 2A). Notably, nearly 40% of the 200 samples were profiled with some but not all omics types; such a missing value scenario is not uncommon in large cohort studies, and MOFA is designed to cope with it (Materials and Methods; Appendix Fig S1). MOFA was configured to combine different likelihood models in order to accommodate the combination of continuous and discrete data types in this study.

MOFA identified 10 factors (minimum explained variance 2% in at least one data type; Materials and Methods). These were robust to algorithm initialization as well as subsampling of the data (Appendix Figs S6 and S7). The factors were largely orthogonal, capturing independent sources of variation (Appendix Fig S6). Among these, Factors 1 and 2 were active in most assays, indicating broad roles in multiple molecular layers (Fig 2B). In contrast, other factors such as Factor 3 or Factor 5 were specific to two data modalities, and Factor 4 was active in a single data modality only. Cumulatively, the 10 factors explained 41% of variation in the drug response data, 38% in the mRNA data, 24% in the DNA methylation data and 24% in the mutation data (Fig 2C).

We also trained MOFA when excluding individual data modalities to probe their redundancy, finding that factors that were active

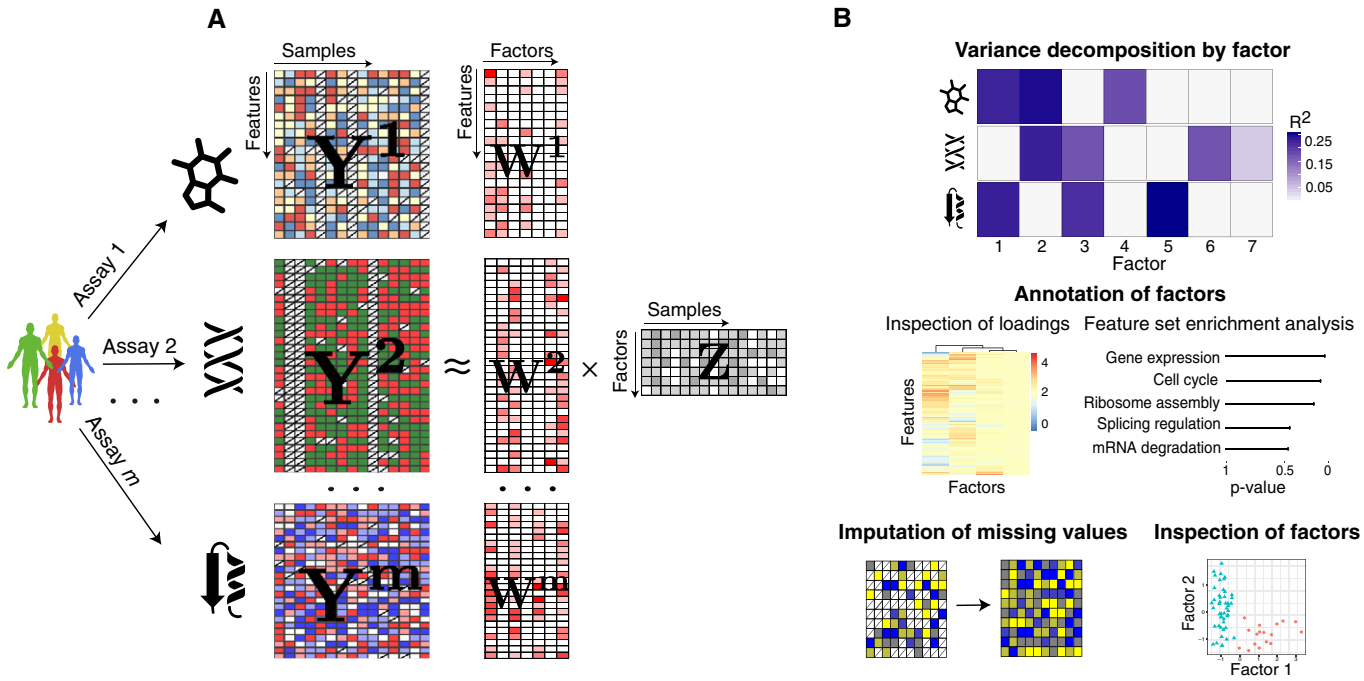


Figure 1. Multi-Omics Factor Analysis: model overview and downstream analyses.

A Model overview: MOFA takes M data matrices as input ($\mathbf{Y}^1,\ldots,\mathbf{Y}^M$), one or more from each data modality, with co-occurrent samples but features that are not necessarily related and that can differ in numbers. MOFA decomposes these matrices into a matrix of factors ($\mathbf{Z}$) for each sample and *M* weight matrices, one for each data modality ($\mathbf{W}^1,\ldots,\mathbf{W}^M$). White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denotes missing values.

B The fitted MOFA model can be queried for different downstream analyses, including (i) variance decomposition, assessing the proportion of variance explained by each factor in each data modality, (ii) semi-automated factor annotation based on the inspection of loadings and gene set enrichment analysis, (iii) visualization of the samples in the factor space and (iv) imputation of missing values, including missing assays.

in multiple data modalities could still be recovered, while the identification of others was dependent on a specific data type (Appendix Fig S8). In comparison with GFA (Leppäaho & Kaski, 2017) and iCluster (Mo *et al*, 2013), MOFA was more consistent in identifying factors across multiple model instances (Appendix Fig S9).

MOFA identifies important clinical markers in CLL and reveals an underappreciated axis of variation attributed to oxidative stress

As part of the downstream pipeline, MOFA provides different strategies to use the loadings of the features on each factor to identify their aetiology (Fig 1B). For example, based on the top weights in the mutation data, Factor 1 was aligned with the somatic mutation status of the immunoglobulin heavy-chain variable region gene (IGHV), while Factor 2 aligned with trisomy of chromosome 12 (Fig 2D and E). Thus, MOFA correctly identified two major axes of molecular disease heterogeneity and aligned them with two of the most important clinical markers in CLL (Zenz *et al*, 2010; Fabbri & Dalla-Favera, 2016; Fig 2D and E).

IGHV status, the marker associated with Factor 1, is a surrogate of the differentiation state of the tumour's cell of origin and the level of activation of the B-cell receptor. While in clinical practice this axis of variation is generally considered binary (Fabbri & Dalla-Favera, 2016), our results indicate a more complex substructure (Fig 3A, Appendix Fig S10). At the current resolution, this factor was consistent with three subgroup models such as proposed by Oakes *et al* (2016) and Queiros *et al* (2015) (Appendix Fig S11), although there is suggestive evidence for an underlying continuum. MOFA connected this factor to multiple molecular layers (Appendix Figs S12 and S13), including changes in the expression of genes previously linked to IGHV status (Vasconcelos *et al*, 2005; Maloum *et al*, 2009; Trojani *et al*, 2012; Morabito *et al*, 2015; Plesingerova *et al*, 2017; Fig 3B and C) and with drugs that target kinases in or downstream of the B-cell receptor pathway (Fig 3D and E).

Despite their clinical importance, the IGHV and the trisomy 12 factors accounted for < 20% of the variance explained by MOFA, suggesting the existence of other sources of heterogeneity. One example is Factor 5, which was active in the mRNA and drug response data. Analysis of the weights in the mRNA revealed that this factor tagged a set of genes enriched for oxidative stress and senescence pathways (Figs 2F and EV2A), with the top weights corresponding to heat-shock proteins (HSPs; Fig EV2B and C), genes that are essential for protein folding and are up-regulated upon stress conditions (Srivastava, 2002; Åkerfelt *et al*, 2010). Although genes in HSP pathways are up-regulated in some cancers and have known roles in tumour cell survival (Trachootham *et al*, 2009), thus far this gene family has received little attention in the context of CLL. Consistent with

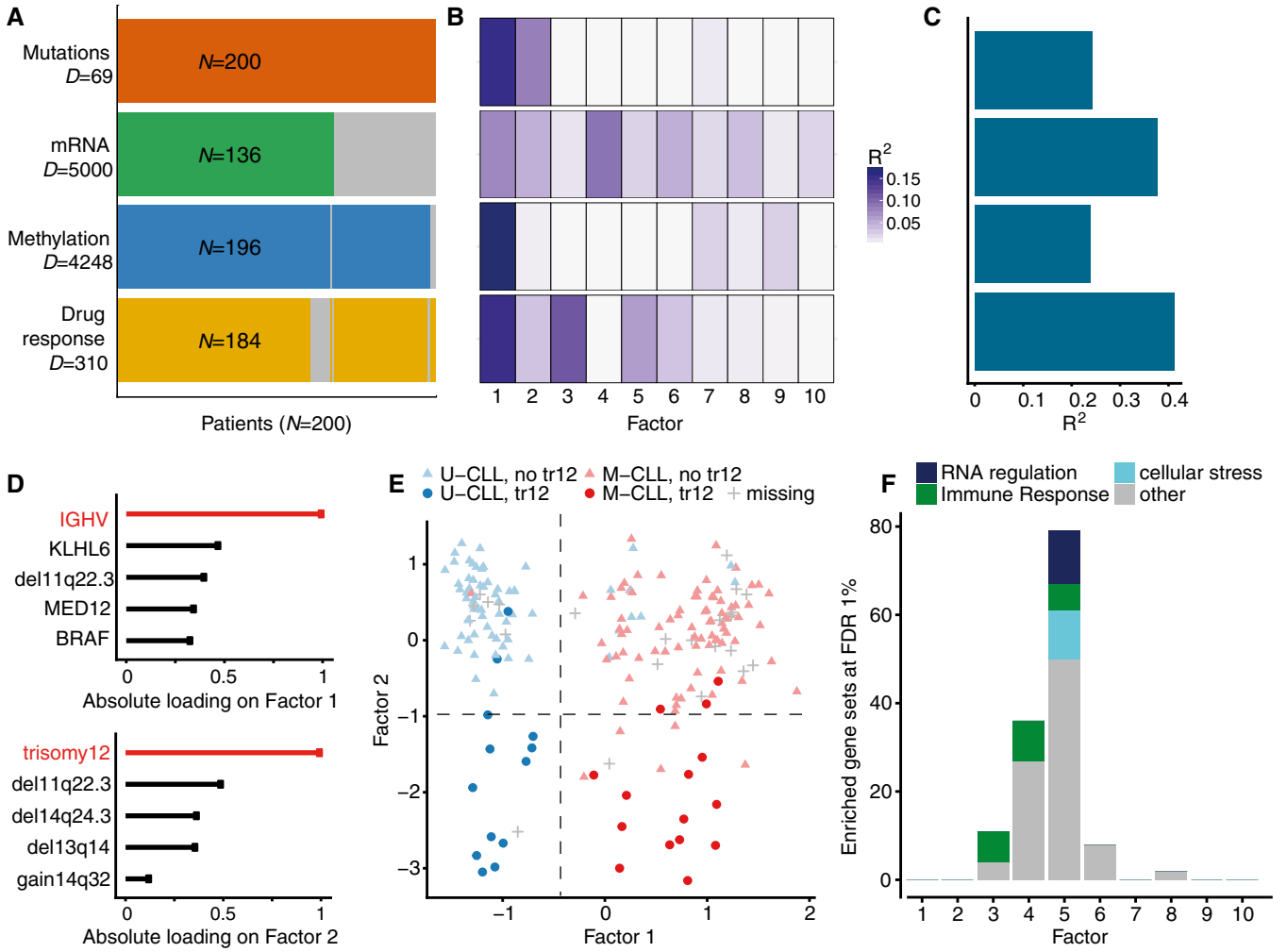


Figure 2. Application of MOFA to a study of chronic lymphocytic leukaemia.

A Study overview and data types. Data modalities are shown in different rows (*D* = number of features) and samples (*N*) in columns, with missing samples shown using grey bars.

B, C (B) Proportion of total variance explained (R^2) by individual factors for each assay and (C) cumulative proportion of total variance explained.

D Absolute loadings of the top features of Factors 1 and 2 in the Mutations data.

E Visualization of samples using Factors 1 and 2. The colours denote the IGHV status of the tumours; symbol shape and colour tone indicate chromosome 12 trisomy status.

F Number of enriched Reactome gene sets per factor based on the gene expression data (FDR < 1%). The colours denote categories of related pathways defined as in Appendix Table S2.

this annotation based on the mRNA data, we observed that the drugs with the strongest weights on Factor 5 were associated with response to oxidative stress, such as target reactive oxygen species (ROS), DNA damage response and apoptosis (Fig EV2D and E).

Factor 4 captured 9% of variation in the mRNA data, and gene set enrichment analysis on the mRNA loadings suggested aetiologies related to immune response pathways and T-cell receptor signalling (Fig 2F), likely due to differences in cell type composition between samples: While the samples are comprised mainly of B cells, Factor 4 revealed a possible contamination with other cell types such as T cells and monocytes (Appendix Fig S14). Factor 3 explained 11% of variation in the drug response data capturing differences in the samples' general level of drug sensitivity (Geeleher *et al*, 2016; Appendix Fig S15).

MOFA identifies outlier samples and accurately imputes missing values

Next, we explored the relationship between inferred factors and clinical annotations, which can be missing, mis-annotated or inaccurate, since they are frequently based on single markers or imperfect surrogates (Westra *et al*, 2011). Since IGHV status is the major biomarker impacting on clinical care, we assessed the consistency between the inferred continuous Factor 1 and this binary marker. For 176 out of 200 patients, the MOFA factor was in agreement with the clinical IGHV status, and MOFA further allowed for classifying 12 patients that lacked clinically measured IGHV status (Fig EV3A and B). Interestingly, MOFA assigned 12 patients to a different group than suggested by their clinical IGHV label. Upon inspection

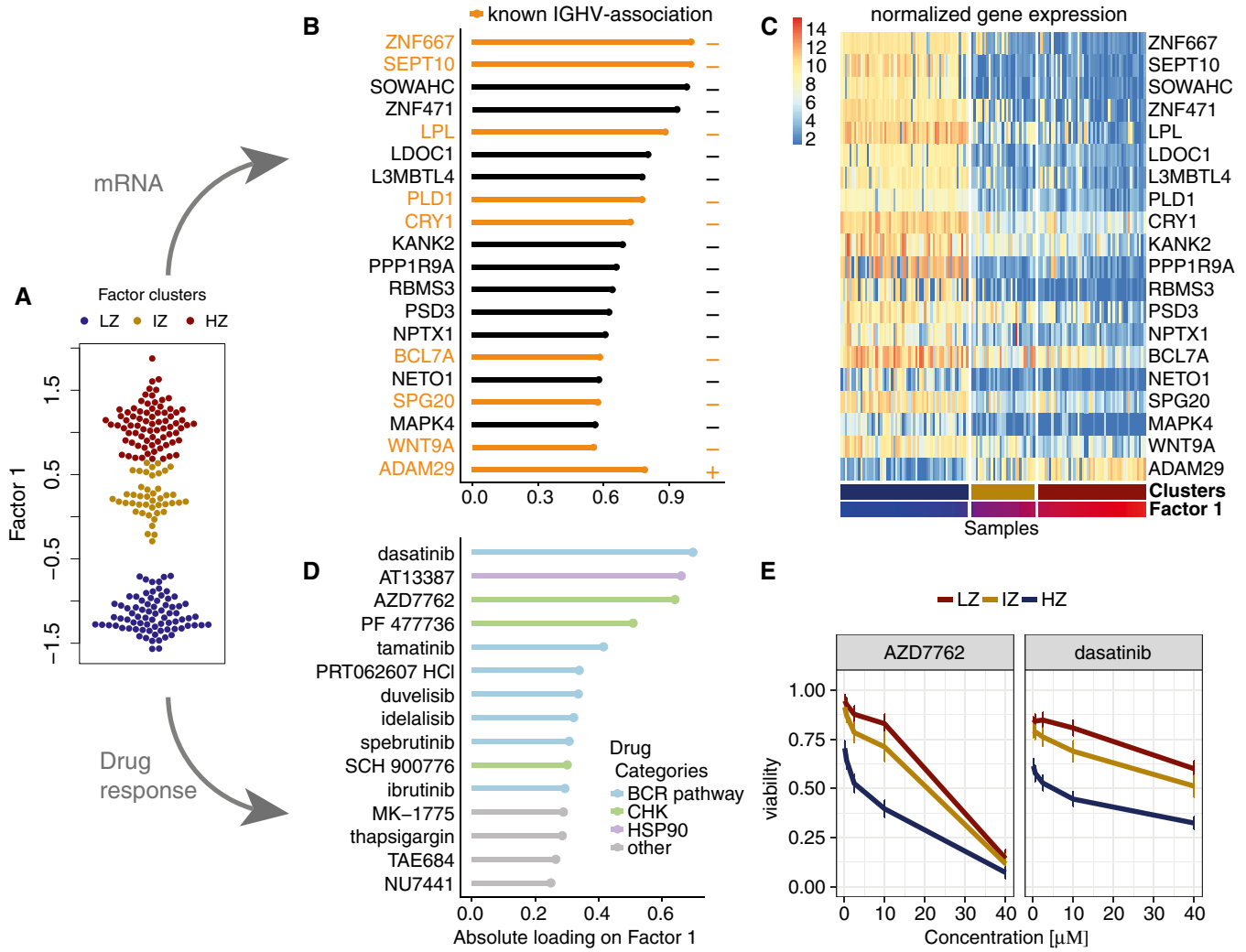


Figure 3. Characterization of the inferred factor associated with the differentiation state of the cell of origin.

A Beeswarm plot with Factor 1 values for each sample with colours corresponding to three groups found by 3-means clustering with low factor values (LZ), intermediate factor values (IZ) and high factor values (HZ).

B Absolute loadings for the genes with the largest absolute weights in the mRNA data. Plus or minus symbols on the right indicate the sign of the loading. Genes highlighted in orange were previously described as prognostic markers in CLL and associated with IGHV status (Vasconcelos *et al*, 2005; Maloum *et al*, 2009; Trojani *et al*, 2012; Morabito *et al*, 2015; Plesingerova *et al*, 2017).

C Heatmap of gene expression values for genes with the largest weights as in (B).

D Absolute loadings of the drugs with the largest weights, annotated by target category.

E Drug response curves for two of the drugs with top weights, stratified by the clusters as in (A).

of the underlying molecular data, nine of these cases showed intermediate molecular signatures, suggesting that they are borderline cases that are not well captured by the binary classification; the remaining three cases were clearly discordant (Fig EV3C and D). Additional independent drug response assays as well as whole exome sequencing data confirmed that these cases are outliers within their IGHV group (Fig EV3E and F).

As incomplete data is a common problem in studies that combine multiple high-throughput assays, we assessed the ability of MOFA to fill in missing values within assays as well as when entire data modalities are missing for some of the samples. For both imputation tasks, MOFA yielded more accurate predictions than other established imputation strategies, including imputation by feature-wise mean, SoftImpute (Mazumder *et al*, 2010) and a k-nearest neighbour method (Troyanskaya *et al*, 2001; Fig EV4, Appendix Fig S16), and MOFA was more robust than GFA, especially in the case of missing assays (Appendix Fig S17).

Latent factors inferred by MOFA are predictive of clinical outcomes

Finally, we explored the utility of the latent factors inferred by MOFA as predictors in models of clinical outcomes. Three of the 10 factors identified by MOFA were significantly associated with time to next treatment (Cox regression, Materials and Methods, FDR < 1%, Fig 4A and B): Factor 1, related to the B-cell of origin,

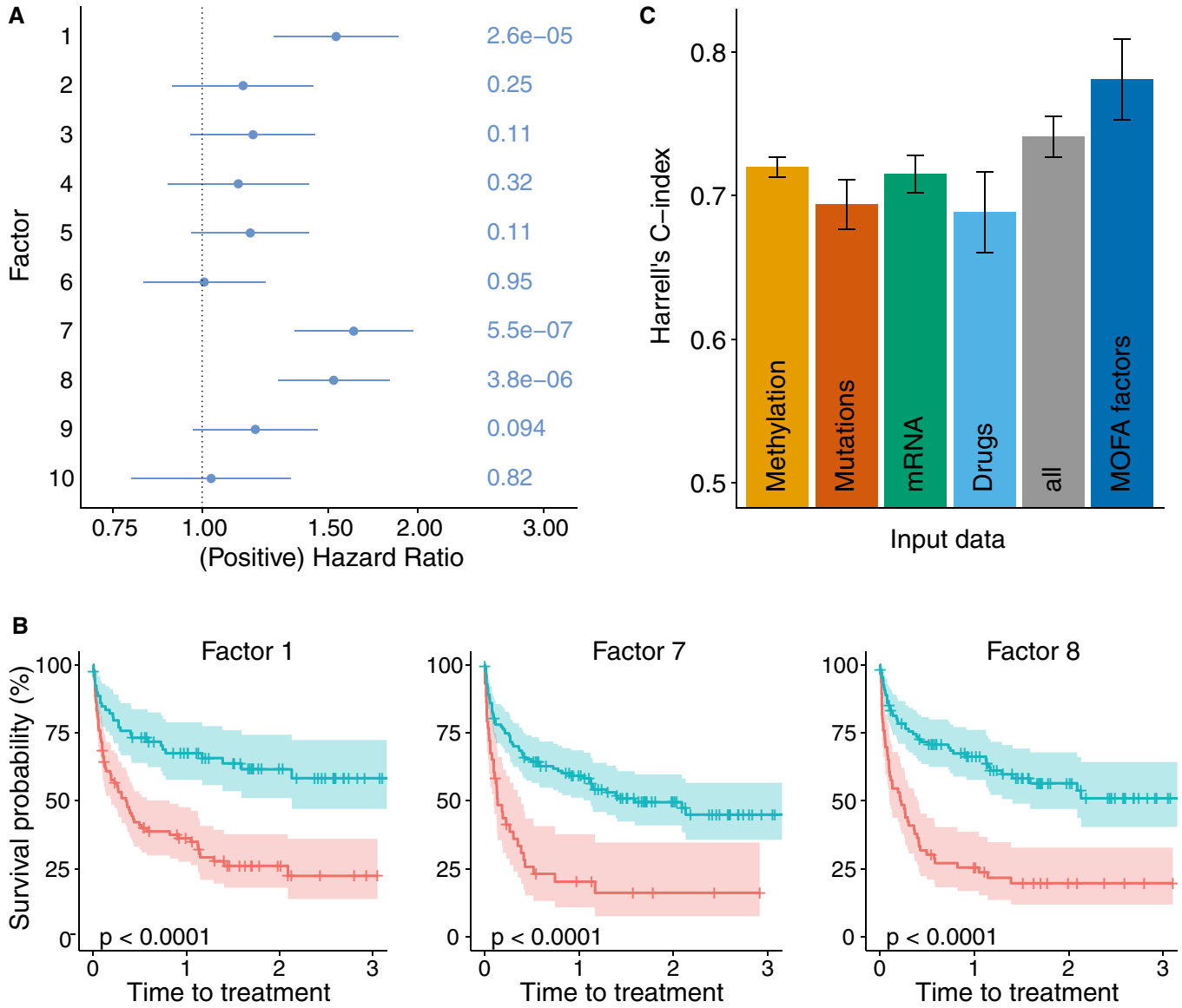


Figure 4. Relationship between clinical data and latent factors.

A Association of MOFA factors to time to next treatment using a univariate Cox regression with $N = 174$ samples (96 of which are uncensored cases) and P-values based on the Wald statistic. Error bars denote 95% confidence intervals. Numbers on the right denote P-values for each predictor.

B Kaplan–Meier plots measuring time to next treatment for the individual MOFA factors. The cut-points on each factor were chosen using maximally selected rank statistics (Hothorn & Lausen, 2003), and P-values were calculated using a log-rank test on the resulting groups.

C Prediction accuracy of time to treatment for $N = 174$ patients using multivariate Cox regression trained using the 10 factors derived using MOFA, as well using the first 10 components obtained from PCA applied to the corresponding single data modalities and the full data set (assessed on hold-out data). Shown are average values of Harrell's C-index from fivefold cross-validation. Error bars denote standard error of the mean.

and two Factors, 7 and 8, associated with chemo-immunotherapy treatment prior to sample collection ($P < 0.01$, t-test). In particular, Factor 7 captures del17p and TP53 mutations as well as differences in methylation patterns of oncogenes (Garg *et al*, 2014; Fluhr *et al*, 2016; Appendix Fig S18), while Factor 8 is associated with WNT signalling (Appendix Fig S19).

We also assessed the prediction performance when combining the 10 MOFA factors in a multivariate Cox regression model. Notably, this model yielded higher prediction accuracy than models using components derived from conventional PCA (Fig 4C), individual molecular features (Appendix Fig S20) or MOFA factors derived from only a subset of the available data modalities (Appendix Fig S8B and D; assessed using cross-validation, Materials and Methods). The predictive value of MOFA factors was similar to clinical covariates (such as lymphocyte doubling time) that are used to guide treatment decisions (Appendix Fig S21).

In an application to single cell data MOFA reveals coordinated changes between the transcriptome and the epigenome along a differentiation trajectory

As multi-omics approaches are also beginning to emerge in single-cell biology (Macaulay *et al*, 2015; Angermueller *et al*, 2016; Guo *et al*, 2017; Clark *et al*, 2018; Colomé-Tatché & Theis, 2018), we investigated the potential of MOFA to disentangle the heterogeneity observed in such studies. We applied MOFA to a data set of 87 mouse embryonic stem cells (mESCs), comprising of 16 cells cultured in "2i" media, which induces a naive pluripotency state, and 71 serum-grown cells, which commits cells to a primed pluripotency state poised for cellular differentiation (Angermueller *et al*, 2016). All cells were profiled using single-cell methylation and transcriptome sequencing, which provides parallel information of these two molecular layers (Fig 5A). We applied MOFA to disentangle the observed heterogeneity in the transcriptome and the CpG methylation at three different genomic contexts: promoters, CpG islands and enhancers.

MOFA identified three major factors driving cell–cell heterogeneity (minimum explained variance of 2%, Materials and Methods): while Factor 1 is shared across all data modalities (7% variance explained in the RNA data and between 53 and 72% in the methylation data sets), Factors 2 and 3 are active primarily in the RNA data

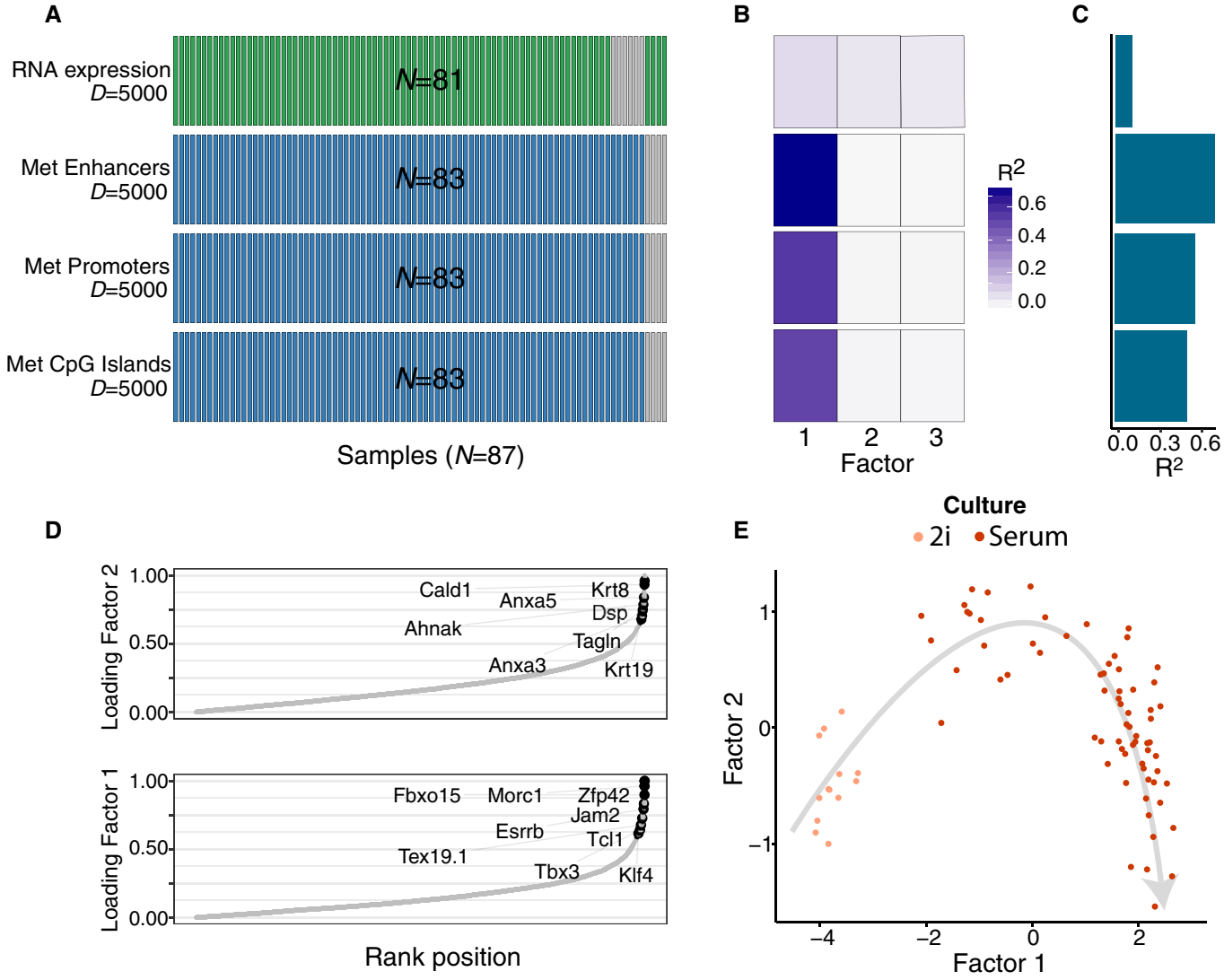


Figure 5. Application of MOFA to a single-cell multi-omics study.

A Study overview and data types. Data modalities are shown in different rows (*D* = number of features) and samples (*N*) in columns, with missing samples shown using grey bars.

B, C (B) Fraction of the variance explained (R^2) by individual factors for each data modality and (C) cumulative proportion of variance explained.

D Absolute loadings of Factor 1 (bottom) and Factor 2 (top) in the mRNA data. Labelled genes in Factor 1 are known markers of pluripotency (Mohammed *et al*, 2017) and genes labelled in Factor 2 are known differentiation markers (Fuchs, 1988).

E Scatterplot of Factors 1 and 2. Colours denote culture conditions. The grey arrow illustrates the differentiation trajectory from naive pluripotent cells via primed pluripotent cells to differentiated cells.

(Fig 5B and C). Gene loadings revealed that Factor 1 captured the cells' transition from naïve to primed pluripotent states, pinpointing pluripotency markers such as Rex1/Zpf42, Tbx3, Fbxo15 and Essrb (Mohammed *et al*, 2017; Figs 5D and EV5A). MOFA connected these transcriptomic changes to coordinated changes in the genome-wide DNA methylation rate across all genomic contexts (Fig EV5B), as previously described both *in vitro* (Angermueller *et al*, 2016) and *in vivo* (Auclair *et al*, 2014). Factor 2 captured a second axis of differentiation from the primed pluripotency state to a differentiated state with highest RNA loadings for known differentiation markers such as keratins and annexins (Fuchs, 1988; Figs 5D and EV5C). Finally, Factor 3 captured the cellular detection rate, a known technical covariate associated with cell quality and mRNA content (Finak *et al*, 2015; Appendix Fig S22).

Jointly, Factors 1 and 2 captured the entire differentiation trajectory from naive pluripotent cells via primed pluripotent cells to differentiated cells (Fig 5E), illustrating the importance of learning continuous latent factors rather than discrete sample assignments. Multi-omics clustering algorithms such as SNF (Wang *et al*, 2014) or iCluster (Shen *et al*, 2009; Mo *et al*, 2013) were only capable of distinguishing cellular subpopulations, but not of recovering continuous processes such as cell differentiation (Appendix Fig S23).

Discussion

Multi-Omics Factor Analysis (MOFA) is an unsupervised method for decomposing the sources of heterogeneity in multi-omics data sets. We applied MOFA to high-dimensional and incomplete multi-omics profiles collected from patient-derived tumour samples and to a single-cell study of mESCs.

First, in the CLL study, we demonstrated that our method is able to identify major drivers of variation in a clinically and biologically heterogeneous disease. Most notably, our model identified previously known clinical markers as well as novel putative molecular drivers of heterogeneity, some of which were predictive of clinical outcome. Additionally, since MOFA factors capture variations of multiple features and data modalities, inferred factors can help to mitigate assay noise, thereby increasing the sensitivity for identifying molecular signatures compared to using individual features or assays. Our results also demonstrate that MOFA can leverage information from multiple omics layers to accurately impute missing values from sparse profiling data sets and guide the detection of outliers, e.g. due to mislabelled samples or sample swaps.

In a second application, we used MOFA for the analysis of single-cell multi-omics data. This use case illustrates the advantage of learning continuous factors, rather than discrete groups, enabling MOFA to recover a differentiation trajectory by combining information from two sparsely profiled molecular layers.

While applications of factor models for integrating different data types were reported previously (Lanckriet *et al*, 2004; Shen *et al*, 2009; Akavia *et al*, 2010; Mo *et al*, 2013), MOFA provides unique features (Materials and Methods, Appendix Table S3) that enable the interpretable reconstruction of the underlying factors and accommodating different data types as well as different patterns of missing data. MOFA is available as open-source software and includes semi-automated analysis pipelines allowing for in-depth characterizations of inferred factors. Taken together, this will foster the accessibility of interpretable factor models for a wide range of multi-omics studies.

Although we have addressed important challenges for multi-omics applications, MOFA is not free of limitations. The model is linear, which means that it can miss strongly non-linear relationships between features within and across assays (Buettner & Theis, 2012). Non-linear extensions of MOFA may address this, although, as with any models in high-dimensional spaces, there will be trade-offs between model complexity, computational efficiency and interpretability (preprint: Damianou *et al*, 2016). A related area of work is to incorporate prior information on the relationships between individual features. For example, future extensions could make use of pathway databases within each omic type (Buettner *et al*, 2017) or priors that reflect relationships given by the "dogma of molecular biology". In addition, new likelihoods and noise models could expand the value of MOFA in data sets with specific statistical properties that hamper the application of traditional statistical methods, including zero-inflated data (i.e. scRNA-Seq; Pierson & Yau, 2015) or binomial distributed data (i.e. splicing events; Huang & Sanguinetti, 2017). Finally, while here we focus our attention on the point estimates of inferred factors, future extensions could attempt a more comprehensive Bayesian treatment that propagates evidence strength and estimation uncertainties to diagnostics and downstream analyses.

Materials and Methods

Multi-Omics Factor Analysis model

Starting from M data matrices $\mathbf{Y}^1,..,\mathbf{Y}^M$ of dimensions $N \times D_m$, where N is the number of samples and D_m the number of features in data matrix m, MOFA decomposes these matrices as

$$\mathbf{Y^m} = \mathbf{ZW^{mT}} + \boldsymbol{\varepsilon^m} \quad \boldsymbol{m} = \boldsymbol{1},\ldots,\boldsymbol{M}. \tag{1}$$

Here, $\mathbf{Z}$ denotes the factor matrix (common for all data matrices) and $\mathbf{W}^{\boldsymbol{m}}$ denotes the weight matrices for each data matrix m (also referred to as view m in the following). $\boldsymbol{\varepsilon^m}$ denotes the view-specific residual noise term, with its form depending on the specifics of the data type (see Noise model).

The model is formulated in a probabilistic Bayesian framework, where we place prior distributions on all unobserved variables of the model (see plate diagram in Appendix Fig S24), i.e. the factors $\mathbf{Z}$, the weight matrices $\mathbf{W}^{\boldsymbol{m}}$ and the parameters of the residual noise term. In particular, we use a standard normal prior for the factors $\mathbf{Z}$ and employ sparsity priors for the weight matrices (see next section).

Model regularization

An appropriate regularization of the weight matrices is essential for the model's ability to disentangle variation across data sets and to yield interpretable factors. MOFA uses a two-level regularization: the first level encourages view- and factor-wise sparsity, thereby allowing to directly identify which factor is active in which view. The second level encourages feature-wise sparsity, thereby typically resulting in a small number of features with active weights. To

encode these sparsity levels, we combine an Automatic Relevance Determination (ARD) prior for the first type of the sparsity with a spike-and-slab prior for the second. For amenable inference, we model the spike-and-slab prior by parameterizing the weights as a product of a Bernoulli distributed random variable and a normally distributed random variable: $W = S\widehat{W}$, where $s_{dk}^m \sim \text{Ber}(\theta_k^m)$ and $\widehat{W}_{dk}^m \sim N(0, 1/\alpha_k^m)$. To automatically learn the appropriate level of regularization for each factor and view, we use uninformative conjugate prior on α_k^m, which controls the strength of factor k in view m, and on θ_k^m, which determines the feature-wise sparsity level of factor k in view m (see Appendix Supplementary Methods, Section 2 for details).

Noise model

MOFA supports the combination of different noise models to integrate diverse data types, including continuous, binary and count data. A standard noise model for continuous data is the Gaussian noise model assuming iid heteroscedastic residuals $\boldsymbol{\varepsilon}^{\mathbf{m}}$, i.e. $\varepsilon_{nd}^m \sim N(0, 1/\tau_d^m)$, with Gamma prior on the precision parameters τ_d^m. MOFA further supports noise models for binary and count data that are not appropriately modelled using a Gaussian likelihood. In the current version, MOFA models count data using a Poisson model and binary data by using a Bernoulli model. Here, the model likelihood is given by $y_{nd}^m \sim \text{Poi}\left(\lambda\left(Z_{n:}w_{d:}^T\right)\right)$ and $y_{nd}^m \sim \text{Ber}\left(\sigma\left(Z_{n:}w_{d:}^T\right)\right)$, respectively, where $\lambda(x) = \log(1 + e^x)$ and σ denotes the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$.

Parameter inference

For scalability, we make use of a variational Bayesian framework, which is essentially a mean field approximation for approximate inference (Blei *et al*, 2017). The key idea is to approximate the intractable posterior distribution using a simpler class of distributions by minimizing the Kullback–Leibler divergence to the exact posterior, or equivalently maximizing the evidence lower bound (ELBO). Convergence of the algorithm can be monitored based on the ELBO. An overview of variational inference and details on the specific implementation for MOFA can be found in Appendix Supplementary Methods, Section 3. To enable an efficient inference for non-Gaussian likelihoods, we employ variational lower bounds on the likelihood (Jaakkola & Jordan, 2000; Seeger & Bouchard, 2012; see Appendix Supplementary Methods, Section 4).

Model training and selection

An important part of the training is the determination of the number of factors. Factors are automatically inactivated by the ARD prior of the model as described in Model regularization. In practice, factors are pruned during training using a minimum fraction of variance explained threshold that needs to be specified by the user. Alternatively, the user can fix the number of factors and the minimum variance criterion is ignored. In the analyses presented, we initialized the models with $K = 25$ factors and they were pruned during training using a threshold of variance explained of 2%. For details on the implementation as well as practical considerations for training and choice of the threshold parameter, refer to Appendix Supplementary Methods, Section 5.

While the inferred factors are robust under different initializations (e.g. Appendix Fig S6C and D), the optimization landscape is non-convex, and hence, the algorithm is not guaranteed to converge to a global optimum. Results presented here are based on 10–25 random restarts, selecting the model with the highest ELBO (e.g. Appendix Fig S6B).

Downstream analysis for factor interpretation and annotation

As part of MOFA, we provide the R package *MOFAtools*, which provides a semi-automated pipeline for the characterization and interpretation of the latent factors. In all downstream analyses, we use the expectations of the model variables under the posterior distributions inferred by the variational framework.

The first step, after a model has been trained, is to disentangle the variation explained by each factor in each view. To this end, we compute the fraction of the variance explained (R^2) by factor k in view m as

$$R_{m,k}^2 = 1 - \left(\sum_{n,d} y_{nd}^m - z_{nk}w_{kd}^m - \mu_d^m\right)^2 / \left(\sum_{n,d} y_{nd}^m - \mu_d^m\right)^2$$

as well as the fraction of variance explained per view taking into account all factors

$$R_m^2 = 1 - \left(\sum_{n,d} y_{nd}^m - \sum_k z_{nk}w_{kd}^m - \mu_d^m\right)^2 / \left(\sum_{n,d} y_{nd}^m - \mu_d^m\right)^2$$

Here, μ_d^m denotes the feature-wise mean. Subsequently, each factor is characterized by three complementary analyses:

1. *Ordination of the samples in factor space:* Visualize a low-dimensional representation of the main drivers of sample heterogeneity.
2. *Inspection of top features with largest weight:* The loadings can give insights into the biological process underlying the heterogeneity captured by a latent factor. Due to scale differences between assays, the weights of different views are not directly comparable. For simplicity, we scale each weight vector by its absolute value.
3. *Feature set enrichment analysis:* Combine the signal from functionally related sets of features (e.g. gene sets) to derive a feature set-based annotation. By default, we use a parametric *t*-test comparing the means of the foreground set (the weights of features that belong to a set G) and the background set (the weights of features that do not belong to the set G), similar to the approach described in Frost *et al* (2015).

Relationship to existing methods

MOFA builds upon the statistical framework of group Factor Analysis (Virtanen *et al*, 2012; Khan *et al*, 2014; Klami *et al*, 2015; Bunte *et al*, 2016; Zhao *et al*, 2016; Leppäaho & Kaski, 2017) and is in part also related to the iCluster methods (Shen *et al*, 2009; Mo *et al*, 2013) as shown in Appendix Table S3. Here, we describe these connections in further detail.

iCluster

In contrast to MOFA, iCluster uses in each view the same extent of regularization for all factors, which may be sufficient for the purpose of clustering (the primary application of iCluster); however,

it results in a reduced ability for distinguishing factors that drive variation in distinct subsets of views (Appendix Fig S5). Additionally, unlike MOFA and GFA, iCluster does not handle missing values and is computationally demanding (Fig EV1), as it requires re-fitting the model for a large range of different penalty parameters and choices of the model dimension.

Group Factor Analysis

While the underlying model of MOFA is closely connected to the most recent GFA implementation (Leppäaho & Kaski, 2017), GFA is restricted to Gaussian observation noise. In terms of the algorithmic implementation, MOFA uses an additional "burn-in period" during training during which the sparsity constraints are deactivated to avoid early splitting of factors and actively drops factors below a predefined variance threshold (see Model training and selection). In contrast, GFA directly uses sparsity constraints throughout training and also maintains factors that have near-zero relevance. In terms of inference, MOFA is implemented using a variational approximate Bayesian inference, whereas GFA is based on a Gibbs sampler. In terms of computational scalability (Fig EV1), both methods are linear in the model's parameters, although GFA is computationally more expensive in absolute terms. This difference is particularly pronounced for data sets with missing data. This, together with the inability to deactivate factors during inference (Appendix Fig S4), renders GFA considerably slower in applications to real data.

Details on the simulation studies

Model validation

To validate MOFA, we simulated data from the generative model for a varying number of views (M = 1,3,...,21), features (D = 100,500,...,10,000), factors (K = 5,10,...,60), missing values (from 0 to 90%) as well as for non-Gaussian likelihoods (Poisson, Bernoulli; see Appendix Table S1 for simulation parameters). We assessed the ability of MOFA to recover the true simulated number of factors in the different settings, where we considered 10 repeat experiments for every configuration. All trials were started with a high number of factors (K = 100), and inactive factors were pruned as described in Model training and selection.

Model comparison

To compare MOFA with to GFA, we simulated data from the underlying generative model with K_{true} = 10 factors, M = 3 views, N = 100 samples, D = 5,000 features each and 5% missing values (missing at random). For each of the three views, we used a different likelihood model: continuous data were simulated with a Gaussian distribution, binary data with a Bernoulli distribution and count data with a Poisson distribution. Except for the non-Gaussian likelihood extension, both methods share the same underlying generative model, thus allowing for a meaningful comparison. We fit ten realizations of the MOFA and GFA models with K_{initial} = 20 factors and let the method determine the most likely number factors. To assess scalability, we considered the same base parameter settings, varying one of the simulation parameters at a time (number of factors K, number of features D, number of samples N and number of views M, all Gaussian). To assess the ability to reconstruct factor activity patterns, we simulated data from the generative model for K_{true} = 10 and K_{true} = 15 factors (M, N, D as before, no missing values, only Gaussian views), where factors were set to either active or inactive in a specific view by sampling the parameter α_k^m from $\{1,10^3\}$. Appendix Table S1 shows in more detail the simulation parameters used in each setting.

Details on the CLL analysis

Data processing

The data were taken from (Dietrich *et al*, 2018), where details on the data generation and processing can be found. Briefly, this data set consists of somatic mutations (combination of targeted and whole exome sequencing), RNA expression (RNA-Seq), DNA methylation (Illumina arrays) and *ex vivo* drug response screens (ATP-based CellTiter-Glo assay). For the training of MOFA, we included 62 drug response measurements (excluding NSC 74859 and bortezomib due to bad quality) at five concentrations each (D = 310) with a threshold at 1.1 to remove outliers. Mutations were considered if present in at least three samples (D = 69). Low counts from RNA-Seq data were filtered out and the data were normalized using the *estimateSizeFactors* and *varianceStabilizingTransformation* function of DESeq2 (Love *et al*, 2014). For training, we considered the top D = 5,000 most variable mRNAs after exclusion of genes from the Y chromosome. Methylation data were transformed to M-values, and we extracted the top 1% most variable CpG sites excluding sex chromosomes (D = 4,248). We included patients diagnosed with CLL and having data in at least two views into the MOFA model leading to a total of N = 200 samples.

Model training and selection

We trained MOFA using 25 random initializations with a variance threshold of 2% and selected the model with the best fit for downstream analysis (see Model training and selection).

Gene set enrichment analysis

Gene set enrichment analysis was performed based on Reactome gene sets (Fabregat *et al*, 2015) as described above. Resulting P-values were adjusted for multiple testing for each factor using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). Significant enrichments were at a false discovery rate of 1%.

Imputation

To compare imputation performance, we trained MOFA on the subset of samples with all measurements (N = 121) and masked at random either single values or all measurements for randomly selected samples in the drug response. After model training, the masked values were imputed directly from the model equation (1) and the accuracy was assessed in terms of mean squared error on the true (masked) values. For both settings, we fixed the number of factors in MOFA to K = 10. To investigate the dependence on K for imputation and to compare MOFA to GFA, we re-ran the same masking experiments varying K = 1,...,20 (Appendix Fig S17).

Survival analysis

Associations between the inferred factors and clinical covariates were assessed using the patients' time to next treatment as

response variable in a Cox model ($N = 174$ samples with treatment information, 96 of which are uncensored cases). For univariate association tests (as shown in Fig 4A, Appendix Fig S21), we scaled all predictors to ensure comparability of the hazard ratios and we rotated factors, which are rotational invariant, such that their hazard ratio is greater or equal to 1. To investigate the predictive power of different data sets, we used a multivariate Cox model and compared Harrell's C-index of predictions in a stratified fivefold cross-validation scheme. As predictors, we included the top 10 principal components calculated on the data for each single view, a concatenated data set ("all") as well as the 10 MOFA factors. Missing values in a view were set to the feature-wise mean. In a second set of models, we used the complete set of all features in a view with a ridge penalty in the Cox model as implemented in the R package *glmnet*. For the Kaplan–Meier plots, an optimal cut-point on each factor was determined to define the two groups using the maximally selected rank statistics as implemented in the R package *survminer* with *P*-values based on a log-rank test between the resulting groups.

Details on the scMT analysis

The data were obtained from Angermueller *et al* (2016), where details on the data generation and pre-processing can be found. Briefly for each CpG site, we calculated a binary methylation rate from the ratio of methylated read counts to total read counts. RNA expression data were normalized using Lun *et al* (2016). To fit MOFA, we considered the top 5,000 most variable genes with a maximum dropout of 90% and the top 5,000 most variable CpG sites with a minimum coverage of 10% across cells. Model selection was performed as described for the CLL data, and factors were inactivated below a minimum explained variance of 2%. For the clustering analysis using SNF and iCluster, the optimal number of clusters was selected using the BIC criterion.

Data and software availability

- The CLL data were obtained from Dietrich *et al* (2018) and are available at the European Genome–Phenome Archive under accession EGAS00001001746 and data tables as R objects can be downloaded from http://pace.embl.de/. The single-cell data were obtained from Angermueller *et al* (2016) and are available in the Gene Expression Omnibus under accession GSE74535. All data used are contained within the MOFA vignettes and can be downloaded as from https://github.com/bioFAM/MOFA.
- An open-source implementation of MOFA in R and Python is available from https://github.com/bioFAM/MOFA. Code to reproduce all the analyses presented is available at https://github.com/bioFAM/MOFA_analysis.

Expanded View for this article is available online.

Acknowledgements
We thank everybody involved in the generation and analysis of the original CLL study for sharing their data and analysis ahead of publication, especially M. Oleś for providing the associated data package and to J. Lu, J. Hüllein and A. Mock for discussions on CLL biology. The work was supported by the European Union (Horizon 2020 project SOUND) and project PanCanRisk).

Author contributions
RA and BV contributed equally and are listed alphabetically. FB, DA and OS conceived the model. RA, DA and BV implemented the model. TZ, SD and WH designed the CLL study and generated the data. RA and BV performed the analysis. RA, BV, DA, TZ, SD, WH, OS, FB and JCM interpreted the results. RA, BV, OS, WH and FB conceived the project. RA, BV, OS, FB and WH wrote the manuscript. OS, WH, FB and JCM supervised the project.

Conflict of interest
The authors declare that they have no conflict of interest.

References
Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D (2010) An integrated approach to uncover drivers of cancer. *Cell* 143: 1005 – 1017

Åkerfelt M, Morimoto RI, Sistonen L (2010) Heat shock factors: integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol* 11: 545

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 13: 229 – 232

Auclair G, Guibert S, Bender A, Weber M (2014) Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol* 15: 545

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57: 289 – 300

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: a review for statisticians. *J Am Stat Assoc* 112: 859 – 877

Buettner F, Theis FJ (2012) A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* 28: i626 – i632

Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 18: 212

Bunte K, Leppaaho E, Saarinen I, Kaski S (2016) Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics* 32: 2457 – 2463

Cancer Genome Atlas Research Network (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* 169: 1327 – 1341.e1323

Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yan Y, Kundu K, Ecker S (2016) Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* 167: 1398 – 1414.e1324

Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* 9: 781

Colomé-Tatché M, Theis F (2018) Statistical single cell multi-omics integration. *Curr Opin Syst Biol* 7: 54 – 59

Damianou A, Lawrence ND, Ek CH (2016) Multi-view learning as a nonparametric nonlinear inter-battery factor analysis. *arXiv* 1604.04939. https://arxiv.org/abs/1604.04939 [PREPRINT]

Dietrich S, Oleś M, Lu J, Sellner L, Anders S, Velten B, Wu B, Hüllein J, da Silva Liberio M, Walther T (2018) Drug-perturbation-based stratification of blood cancer. *J Clin Invest* 128: 427 – 445

Fabbri G, Dalla-Favera R (2016) The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat Rev Cancer* 16: 145 – 162

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S (2015) The reactome pathway knowledgebase. *Nucleic Acids Res* 44: D481 – D487

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16: 278

Fluhr S, Boerries M, Busch H, Symeonidi A, Witte T, Lipka DB, Mücke O, Nöllke P, Krombholz CF, Niemeyer CM (2016) CREBBP is a target of epigenetic, but not genetic, modification in juvenile myelomonocytic leukemia. *Clin Epigenet* 8: 50

Frost HR, Li Z, Moore JH (2015) Principal component gene set enrichment (PCGSE). *BioData Min* 8: 25

Fuchs E (1988) Keratins as biochemical markers of epithelial differentiation. *Trends Genet* 4: 277 – 281

Garg R, Benedetti LG, Abera MB, Wang H, Abba M, Kazanietz MG (2014) Protein kinase C and cancer: what we know and what we do not. *Oncogene* 33: 5225 – 5237

Geeleher P, Cox NJ, Huang RS (2016) Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biol* 17: 190

Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Porta MG, Jadersten M, Dolatshad H, Verma A, Cross NC, Vyas P, Killick S, Hellstrom-Lindberg E, Cazzola M, Papaemmanuil E, Campbell PJ, Boultwood J (2015) Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun* 6: 5901

GTEx Consortium (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348: 648 – 660

Guo F, Li L, Li J, Wu X, Hu B, Zhu P, Wen L, Tang F (2017) Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res* 27: 967 – 988

Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. *Genome Biol* 18: 83

Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, Marchini J (2016) Tensor decomposition for multiple-tissue gene expression experiments. *Nat Genet* 48: 1094 – 1100

Hothorn T, Lausen B (2003) On the exact distribution of maximally selected rank statistics. *Comput Stat Data Anal* 43: 121 – 137

Huang Y, Sanguinetti G (2017) BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol* 18: 123

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Goncalves E, Barthorpe S, Lightfoot H, Cokelaer T, Greninger P, van Dyk E, Chang H, de Silva H, Heyn H, Deng X, Egan RK, Liu Q, Mironenko T *et al* (2016) A landscape of pharmacogenomic interactions in cancer. *Cell* 166: 740 – 754

Jaakkola TS, Jordan MI (2000) Bayesian parameter estimation via variational methods. *Stat Comput* 10: 25 – 37

Khan SA, Virtanen S, Kallioniemi OP, Wennerberg K, Poso A, Kaski S (2014) Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics* 30: i497 – i504

Kim M, Rai N, Zorraquino V, Tagkopoulos I (2016) Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nat Commun* 7: 13090

Klami A, Virtanen S, Leppaaho E, Kaski S (2015) Group factor analysis. *IEEE Trans Neural Netw Learn Syst* 26: 2136 – 2147

Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) A statistical framework for genomic data fusion. *Bioinformatics* 20: 2626 – 2635

Leppäaho E, Kaski S (2017) GFA: exploratory analysis of multiple data sources with group factor analysis. *J Mach Learn Res* 18: 1 – 5

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Lun AT, Bach K, Marioni JC (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 17: 75

Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, Smith M, Van der Aa N, Banerjee R, Ellis PD, Quail MA, Swerdlow HP, Zernicka-Goetz M, Livesey FJ, Ponting CP, Voet T (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* 12: 519 – 522

Maloum K, Settegrana C, Chapiro E, Cazin B, Lepretre S, Delmer A, Leporrier M, Dreyfus B, Tournilhac O, Mahe B, Nguyen-Khac F, Lesty C, Davi F, Merle-Beral H (2009) IGHV gene mutational status and LPL/ADAM29 gene expression as clinical outcome predictors in CLL patients in remission following treatment with oral fludarabine plus cyclophosphamide. *Ann Hematol* 88: 1215 – 1221

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. *J Mach Learn Res* 11: 2287 – 2322

Meng C, Kuster B, Culhane AC, Gholami AM (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 15: 162

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 17: 628 – 641

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, Kawaler E, Mundt F, Krug K, Tu Z, Lei JT, Gatza ML, Wilkerson M, Perou CM, Yellapantula V, Huang KL *et al* (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534: 55 – 62

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA* 110: 4245 – 4250

Mohammed H, Hernando-Herraez I, Savino A, Scialdone A, Macaulay I, Mulas C, Chandra T, Voet T, Dean W, Nichols J (2017) Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep* 20: 1215 – 1228

Morabito F, Cutrona G, Mosca L, D'Anca M, Matis S, Gentile M, Vigna E, Colombo M, Recchia AG, Bossio S, De Stefano L, Maura F, Manzoni M, Ilariucci F, Consoli U, Vincelli I, Musolino C, Cortelezzi A, Molica S, Ferrarini M *et al* (2015) Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment. *Leuk Res* 39: 840 – 845

Oakes CC, Seifert M, Assenov Y, Gu L, Przekopowitz M, Ruppert AS, Wang Q, Imbusch CD, Serva A, Koser SD, Brocks D, Lipka DB, Bogatyrova O, Weichenhan D, Brors B, Rassenti L, Kipps TJ, Mertens D, Zapatka M, Lichter P *et al* (2016) DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat Genet* 48: 253 – 264

Pierson E, Yau C (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* 16: 241

Plesingerova H, Librova Z, Plevova K, Libra A, Tichy B, Skuhrova Francova H, Vrbacky F, Smolej L, Mayer J, Bryja V, Doubek M, Pospisilova S (2017) COBLL1, LPL and ZAP70 expression defines prognostic subgroups of chronic lymphocytic leukemia patients with high accuracy and correlates with IGHV mutational status. *Leuk Lymphoma* 58: 70 – 79

Queiros AC, Villamor N, Clot G, Martinez-Trillos A, Kulis M, Navarro A, Penas EM, Jayne S, Majid A, Richter J, Bergmann AK, Kolarova J, Royo C, Russinol N,

Castellano G, Pinyol M, Bea S, Salaverria I, Lopez-Guerra M, Colomer D *et al* (2015) A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* 29: 598 – 605

Remes S, Mononen T, Kaski S (2015) Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers. *arXiv* 1512.05610. https://arxiv.org/abs/1512.05610 [PREPRINT]

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 16: 85 – 97

Seeger M, Bouchard G (2012) Fast variational Bayesian inference for non-conjugate matrix factorization models. *Artific Intell Stat* 22: 1012 – 1018

Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25: 2906 – 2912

Singh A, Gautier B, Shannon CP, Rohart F, Vacher M, Tebutt SJ, Le Cao K-A (2018) DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv* https://doi.org/10.1101/067611 [PREPRINT]

Soderholm S, Fu Y, Gaelings L, Belanov S, Yetukuri L, Berlinkov M, Cheltsov AV, Anders S, Aittokallio T, Nyman TA, Matikainen S, Kainov DE (2016) Multi-omics studies towards novel modulators of influenza A virus-host interaction. *Viruses* 8: 269

Srivastava P (2002) Roles of heat-shock proteins in innate and adaptive immunity. *Nat Rev Immunol* 2: 185

Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics* 15: 569 – 583

Trachootham D, Alexandre J, Huang P (2009) Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nat Rev Drug Discovery* 8: 579 – 591

Trojani A, Di Camillo B, Tedeschi A, Lodola M, Montesano S, Ricci F, Vismara E, Greco A, Veronese S, Orlacchio A (2012) Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia. *Cancer Biomarkers* 11: 15 – 28

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17: 520 – 525

Vasconcelos Y, De Vos J, Vallat L, Reme T, Lalanne AI, Wanherdrick K, Michel A, Nguyen-Khac F, Oppezzo P, Magnac C, Maloum K, Ajchenbaum-Cymbalista F, Troussard X, Leporrier M, Klein B, Dighiero G, Davi F, French Cooperative Group on CLL (2005) Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes. *Leukemia* 19: 2002 – 2005

Virtanen S, Klami A, Khan S, Kaski S (2012) Bayesian group factor analysis. *Artific Intell Stat* 22: 1269 – 1277

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 11: 333 – 337

Westra H-J, Jansen RC, Fehrmann RS, te Meerman GJ, Van Heel D, Wijmenga C, Franke L (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27: 2104 – 2111

Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S (2010) From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nat Rev Cancer* 10: 37 – 50

Zhao S, Gao C, Mukherjee S, Engelhardt BE (2016) Bayesian group factor analysis with structured sparsity. *J Mach Learn Res* 17: 1 – 47

GigaScience, 9, 2020, 1–10

doi: 10.1093/gigascience/giaa112
DATA NOTE

DATA NOTE

A molecular map of lung neuroendocrine neoplasms

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

[1]Section of Genetics, International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France; and [2]Section of Mechanisms of Carcinogenesis, International Agency for Research on Cancer (IARC-WHO), 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France

*Correspondence address. Matthieu Foll, Section of Genetics, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France. E-mail: follm@iarc.fr http://orcid.org/0000-0001-9006-8436
†Contributed equally.
‡Jointly supervised.

Abstract

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

Keywords: carcinoids; lung cancer; neuroendocrine neoplasms; rare cancers; genomics; Tumormap; lungNENomics project

Background

Lung neuroendocrine neoplasms (LNENs) are rare understudied diseases with limited therapeutic opportunities. LNENs include poorly differentiated and highly aggressive lung neuroendocrine carcinomas (NECs)—i.e., small-cell lung cancer (SCLC) and large-cell neuroendocrine carcinoma (LCNEC)—as well as well-differentiated and less aggressive lung neuroendocrine tumors (NETs), i.e., typical and atypical carcinoids (WHO classification 2015 [1]). Over the past years several genomic studies have

Received: 21 January 2020; **Revised:** 4 June 2020

investigated the molecular characteristics of these diseases to provide some evidence for more personalized clinical management [2–8]. Although lung NECs and NETs are broadly considered different diseases, several recent studies have suggested that they may share some molecular characteristics [7, 9–12]. However, owing to the rarity of these diseases, the sample sizes of these studies individually are limited, and the integration of independent datasets is not an easy task.

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

Data Description

Recently [7], we performed the first integrative and comparative genomic analysis of LNEN samples from all histological types, based on newly sequenced data: whole-exome sequencing (WES) data (16 samples), whole-genome sequencing (WGS) data (3 samples), RNA-sequencing (RNA-Seq) data (20 samples), and EPIC 850K methylation data (76 samples), as well as publicly available data. These data correspond to the most extensive multi-omic dataset of LNENs, including the first methylation data for LCNEC and the first molecular characterization of the rarest LNEN subtype (atypical carcinoids) [7]. This dataset, which provides the missing pieces for a complete molecular characterization of LNENs, has been deposited at the EMBL-EBI European Genome-phenome Archive (EGA accession No. EGAS00001003699). To facilitate the reuse of the data generated for the previous publication [7], we provide here a complementary data descriptor by outlining the pre-processing and quality control (QC) steps performed on each omic dataset available on EGA.

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

Data quality controls

Fig. 1 provides a schematic view of the pre-processing steps and the associated QC performed for each omic dataset generated by Alcala and colleagues [7]. An overview of the available omics and clinical data for each sample is provided in Supplementary Table 1.

WES and WGS data

WES and WGS were performed, respectively, on 16 and 3 fresh-frozen atypical carcinoids in the Cologne Centre for Genomics and the Centre National de Recherche en Génomique Humaine. For WES, the SeqCap EZ v2 Library capture kit from NimbleGen (44 Mb) and the Illumina HiSeq 2000 machine (Illumina Inc., San Diego, CA, USA) were used for the sequencing. For WGS, the Illumina TruSeq DNA PCR-Free Library Preparation Kit was used for library preparation and the HiSeqX5 platform from Illumina for the sequencing as described in [7]. The sequencing reads from the 16 atypical carcinoids' whole exomes and the 3 carcinoids' whole genomes were processed using the in-house Nextflow [15] workflow available at the IARCbioinfo/alignment-nf [16] GitHub repository, revision No. 9092214665. The pipeline consists in 3 steps: mapping reads to the reference genome (GRCh37), marking duplicates, and sorting reads using bwa v0.7.12-r1044 (BWA, RRID:SCR_010910) [17], samblaster v0.1.22 (samblaster, RRID:SCR_000468) [18], and sambamba v0.5.9 [19], respectively. For WES samples, local realignment using ABRA v0.97b (ABRA, RRID:SCR_003277) [20] was then run.

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

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

RNA-Seq data

RNA-Seq was performed on 20 fresh-frozen atypical samples. The Illumina TruSeq RNA sample preparation Kit was used for library preparation and the Illumina TruSeq PE Cluster Kit v3 and the Illumina TruSeq SBS Kit v3-HS kits were used on an Illumina HiSeq 2000 sequencer. The data generated were processed in 5 steps (Fig. 1, middle panel): (i) read trimming using Trim Galore v0.6.5 (Trim Galore, RRID:SCR_011847) [26], (ii) read mapping to the reference genome (GRCh38, gencode version 33 from bundle CTAT from 6 April 2020 [27]) using STAR v.2.7.3a (STAR, RRID:SCR_015899) [28], (iii) realignment of the reads using ABRA2 v2.22 (ABRA, RRID:SCR_003277) [29], (iv) base quality score recalibration using GATK4 v4.0.5.1 (GATK, RRID:SCR_001876) [30, 31], and (v) gene expression quantification using StringTie v2.1.1 (StringTie, RRID:SCR_016323) [32]. FastQC v.0.11.9 (FastQC, RRID:SCR_014583) [21], RSeQC v3.0.1 (RSeQC, RRID:SCR_005275) [33], and HTSeq v0.12.4 (HTSeq, RRID:SCR_005514) [34] were used to control the raw read quality and assignments, and the results aggregated using MultiQC v1.7 (MultiQC, RRID:SCR_014982) [25]. These steps were performed using our in-house Nextflow [15] pipelines available at the following GitHub repositories: IARCbioinfo/RNAseq-nf [35] release v2.3, IARCbioinfo/abra-nf [36] release v3.0, IARCbioinfo/BQSR-nf [37] release v1.1, and IARCbioinfo/RNAseq-transcript-nf [38] release v2.1.

Fig. 2C shows that the base calls, before trimming, are of good quality because all samples have a mean per base sequence quality score >28 (left panel) and for all samples the most fre-

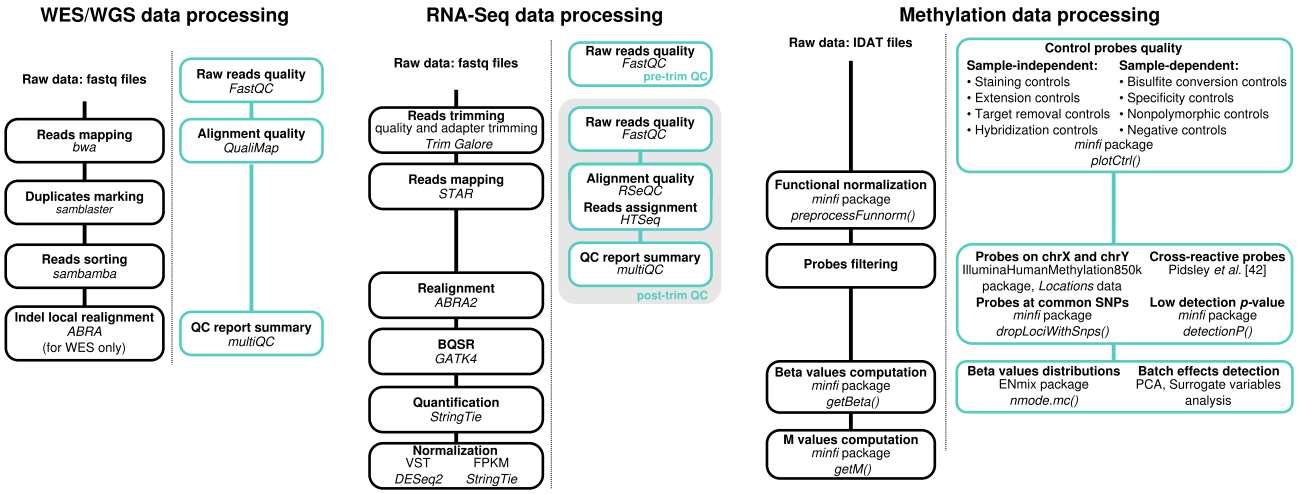


Figure 1: Bioinformatics workflows for data processing and associated quality controls (QC; green boxes). Bioinformatics tools used for the processing of the WES/WGS data, RNA-Seq, and methylation data are represented in the left, middle, and right panels, respectively.

Table 1: General statistics associated with the quality controls of the WES and WGS data

Sample	Sequencing	Median coverage	Total No. reads (M)	>30× (%)	Aligned (%)	GC content (%)	Median insert size	Duplicates (%)
LNEN002	WES	148	113.3	95.5	99.7	53.7	194	13.9
LNEN003	WES	146	110.3	95.8	99.7	53.7	194	13.4
LNEN004	WES	150	115.3	95.4	99.8	54.3	193	13.1
LNEN005	WES	135	103.4	94.7	99.8	54.0	195	12.1
LNEN006	WES	126	93.6	94.6	99.8	53.5	197	12.5
LNEN007	WES	145	116.3	94.4	99.8	54.5	195	14.8
LNEN009	WES	123	98.4	92.9	99.7	54.1	195	12.4
LNEN010	WES	138	104.1	95.0	99.7	53.3	196	13.4
LNEN011	WES	161	125.8	95.8	99.8	54.3	196	14.8
LNEN013	WES	131	99.2	94.3	99.8	53.5	193	13.0
LNEN014	WES	132	102.6	94.0	99.8	54.1	195	13.3
LNEN015	WES	148	111.3	95.7	99.6	54.1	197	10.1
LNEN016	WES	133	98.0	94.3	99.6	54.3	194	9.0
LNEN017	WES	158	116.4	95.9	99.6	54.1	192	8.9
LNEN020	WES	187	144.7	96.6	99.7	53.6	192	14.5
S00716_B	WES	133	99.8	95.4	99.7	52.8	194	14.3
LNEN041	WGS	36	923.5	77.5	98.9	41.0	366	13.3
LNEN042	WGS	41	993.7	88.1	98.8	41.5	388	9.4
LNEN043	WGS	43	1033.1	89.7	99.3	41.6	392	8.8

GC: guanine-cytosine.

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

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

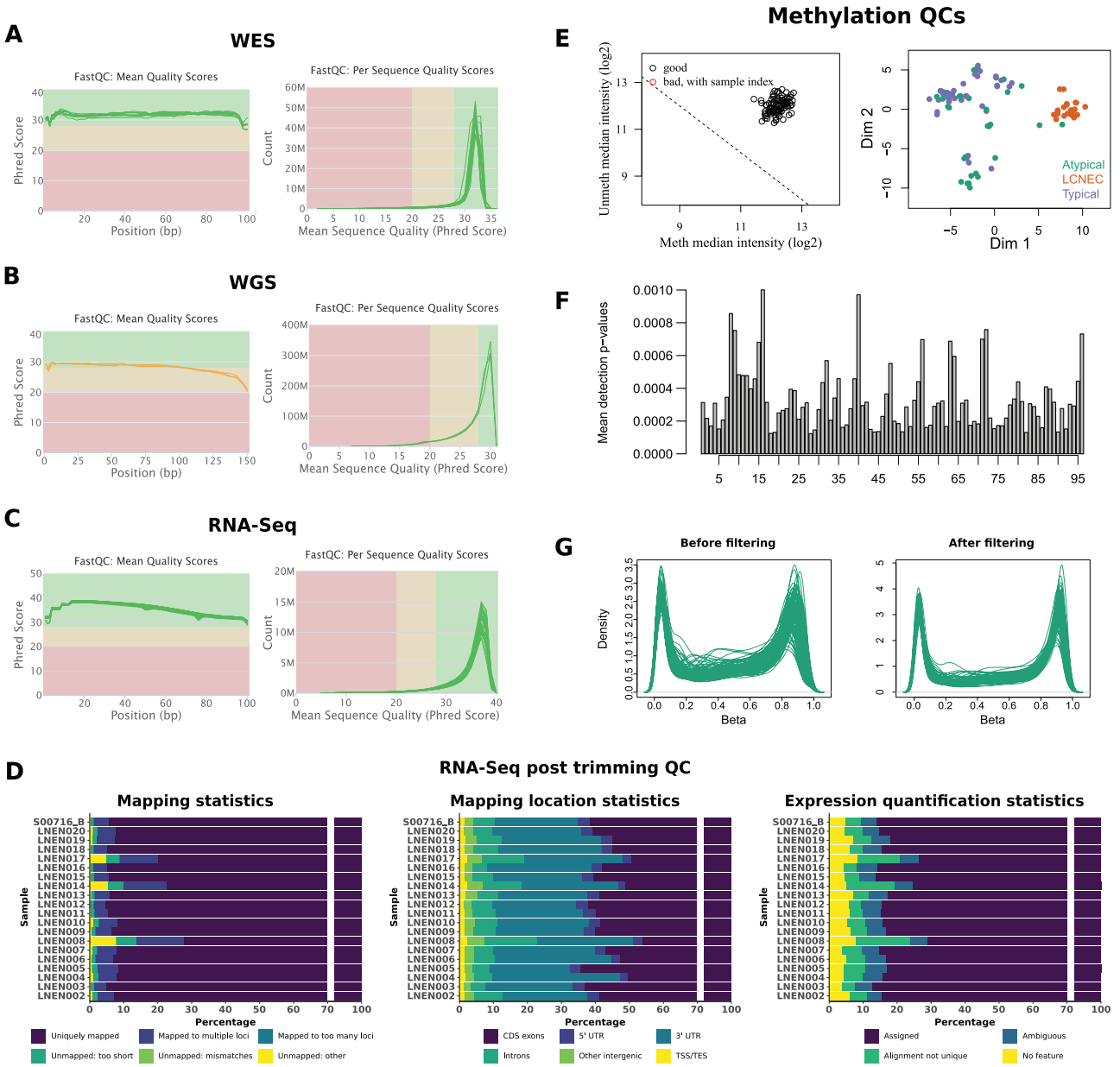


Figure 2: Quality control (QC) performed on each omic dataset. **(A)** Read QC using FastQC for WES data. **(B)** Read QC using FastQC for WGS data. **(C)** Read QC using FastQC for RNA-Seq data. For A, B, and C, the left panels correspond to the sequence quality plots, the x-axis representing the base position in the read and the y-axis the mean quality value; the right panels correspond to the per sequence quality score plots, the x-axis representing the mean quality score and the y-axis the number of reads. **(D)** QC of the RNA-Seq data after trimming. *Left:* Bar plot representing the percentage of reads uniquely mapped ("Uniquely mapped"), mapped to multiple loci ("Mapped to multiple loci" or "Mapped to too many loci" if the number of loci is >10), unmapped because the mapped reads' proportion was too small ("Unmapped: too short"), unmapped because of too many mismatches ("Unmapped: mismatches"), or unmapped for other reasons ("Unmapped: other"). *Middle:* Cumulative bar plot representing the percentages of reads mapped, using RSeQC, at different locations in the genome (exons, introns, 5′ and 3′ untranslated transcribed region [UTR], intergenic regions, TSS, and TES). *Right:* Cumulative bar plot representing the cumulative percentages associated with the different read assignments using HTSeq ("Assigned": reads assigned to 1 gene, "Ambiguous": reads assigned to multiple overlapping genes, "Aligned not unique": reads assigned to multiple non-overlapping genes, "No Feature": reads assigned to none of the features). **(E)** *Left:* Samples' quality based on log median intensities. The x-axis and y-axis correspond to the median of $\log_2$ methylated and unmethylated intensities, respectively. *Right:* Representation of the between-sample similarities based on the 2 first multidimensional scaling dimensions. **(F)** Histogram of the median detection *P*-value for each sample. **(G)** Distribution of the β-values for each sample before and after the filtering step (left and right panel, respectively).

Methylation data

The methylation analyses were performed on the basis of the EPIC 850K methylation arrays and the Infinium EPIC DNA methylation beadchip platform (Illumina) for 33 typical carcinoids, 23 atypical carcinoids, 20 LCNECs, and 19 technical replicates in total. These arrays interrogate >850,000 CpGs and contain internal control probes that can be used to assess the overall efficiency of the sample preparation steps. The raw intensity data (IDAT files) were processed using the R package minfi v.1.24.0 (minfi, RRID:SCR_012830) [40]. Fig. 1 (right panel) provides the packages, functions, and publication used for the data

processing, QC, and filtering steps as implemented in the IARCbioinfo/Methylation_analysis_scripts [41] GitHub repository.

Fig. 2E shows that no outliers were detected: (i) the left panel, representing the median $\log_2$ of the methylated and unmethylated intensities, indicates that all samples cluster together with a log median intensity >11 for both channels, which supports the absence of failed samples; (ii)in the right panel, the multidimensional scaling plot shows that the samples cluster together by histological groups. We used the depectionP function (minfi package), which compares the DNA signal to the background signal based on the negative control probes to provide a detection *P*-value per probe, lower *P*-value indicating reliable CpGs. Fig. 2F represents the mean detection *P*-values per sample and shows that all samples' mean detection *P*-values were <0.01. To correct for the variability identified in the control probes, a normalization step was applied to the raw intensities using the preprocessFunnorm function from minfi.

After between-array normalization, different sets of probes that could generate artifacts were removed successively from the methylation dataset: (i) 19,634 probes on the sex chromosomes, in order to identify differences related to tumors but unrelated to sex chromosomes; (ii) 41,818 cross-reactive probes, which are probes co-hybridizing with multiple CpGs on the genome and not only to the one for which it has been designed [42]; (iii) 10,588 probes associated with common SNPs (present in dbSNP build 137); (iv) 24,363 probes with multi-modal β-value distribution; and (v) 9,697 probes having a detection *P*-value >0.01 in ≥1 sample. Supplementary Table 5 lists the sets of filtered probes. To assess the experimental quality of the assay, the distributions of the β-values were analyzed. As described previously, probes with multi-modal distributions were removed at the filtering step and overall distributions of β-values for each sample before and after filtering were plotted (Fig. 2G). As expected, after filtering all samples showed a bimodal profile, indicative of the good quality of the experiment. No experimental batch effects were identified after functional normalization (see Supplementary Fig. 33 from [7]). Based on all the QCs performed, none of the samples analyzed were identified as outlier. However, 1 sample available on EGA (201414140007_R06C01) was removed from the analyses because it came from a metastatic tumor rather than the primary tumor. Sample metadata are provided in Supplementary Table 6.

Generation of an integrative molecular map

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

Dimensionality reduction using UMAP

UMAP method

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

Biological interpretation of the pan-LNEN TumorMap

Fig. 3 shows the pan-LNEN map available on TumorMap [45] (see "Reuse potential" section below), with colors representing the main molecular subtypes. To evaluate the accuracy of the generated pan-LNEN map we first verified whether it was consistent with the main biological findings from the original studies, in particular whether it represented the molecular subtypes of LNENs previously identified, and their relationship with histological types. We specifically tested whether groups of samples previously described as having discordant molecular and histopathological features were identified in our map. To do so, given a focal molecular subtype and 2 reference histopathological types, we assessed whether samples from the focal molecular subtype were closer to 1 of the 2 references using a 1-sided Wilcoxon test between the Euclidean distances of samples to the centroid of each reference type.

First, the SCLC/LCNEC-like samples [6], which are histological SCLCs presenting the molecular profile of LCNEC, tend to cluster with the LCNECs rather than with the SCLCs (Wilcoxon $P = 6.2 \times 10^{-4}$). Similarly, the LCNEC/SCLC-like samples [6], which are histological LCNECs having the molecular profile of SCLC, tend to cluster with the SCLCs rather than with the LCNECs (Wilcoxon $P = 3.3 \times 10^{-3}$). In 2018, George et al. showed also that LCNEC samples can be subdivided into Type I and Type II molecular groups [6]. We observed that the Type I and Type II LCNECs were closer to each other than to the SCLC/SCLC-like (Wilcoxon $P = 9.9 \times 10^{-14}$) and that SCLC/LCNEC-like samples were closer to Type II than to Type I LCNECs [6] (Wilcoxon $P = 3.9 \times 10^{-3}$). Like the LCNECs, pulmonary carcinoids have been subdivided into molecular groups. Alcala et al. [7] identified 3 clinically relevant molecular clusters, using a multi-omics factor analysis: Carcinoid A1, Carcinoid A2, and Carcinoid B [7]. In the pan-LNEN map generated using UMAP, those 3 clusters are clearly visible (Fig. 3) and, respectively, correspond to the 3 clusters identified in [8] named LC1, LC3, and LC2. Also, in the study from Alcala and colleagues [7], 2 carcinoids that clustered with the carcinoids B (S00118 and S00089) were borderline and located between cluster A1 and B. Similarly, an LCNEC sample and an SCLC sample clustered with the carcinoids A1 [7]. These observations are also visible on the TumorMap representation. Finally, in the same study, a novel entity of carcinoids, named the "supra-carcinoids," was unveiled. These samples were characterized by a morphology similar to that of pulmonary carcinoids but with the molecular features of LCNEC samples. In the pan-LNEN TumorMap, the supra-carcinoids also clustered with the LCNEC samples and were molecularly closer to LCNECs than to SCLCs (Wilcoxon $P = 5 \times 10^{-2}$). We also note that 1 sample from Laddha et al. [8] LC2 cluster (SRR7646258) clusters with LCNEC.

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

Quality control of the UMAP projection

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

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

Preservation of the samples' neighborhood

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

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

By definition, $\overline{SD}'_{k,PCA-5D} = 0$ and $\overline{SD}'_{k,PCA-2D} = 1$. Thus values of $\overline{SD}'_{k,X}$ close to 0 indicate that X preserves k neighborhoods as well as PCA-5D, whereas values close to 1 indicate that X preserves k neighborhoods worse than PCA-5D but as well as PCA-2D, and values >1 indicate that X preserves k neighborhoods worse than PCA-2D and PCA-5D. Note that $\overline{SD}'_{k,X}$ can be negative if X preserves k neighborhoods better than $\overline{SD}_{k,PCA-5D}$. For the UMAP projection, we iterated the computation of $\overline{SD}'_k$ 1,000 times because the algorithm uses a stochastic optimization step to define the projection.

As expected, increasing the n_neighbors UMAP parameter from 15 to 238 leads to a better preservation of the global structure, clearly visible for $k > 30$ (Fig. 4A; mean $\overline{SD}'_{k>30} = 2.855$ and 1.029, respectively), while only marginally reducing the preservation of the local structure for $k < 30$ (mean $\overline{SD}'_{k<30} = -0.076$ and 0.124, respectively), which is approximately the size of the smallest cluster. Globally, the $\overline{SD}'_k$ values over all k levels are lower for an n_neighbors value of 238 than 15 (paired t-test $P = 6.09 \times 10^{-8}$). With n_neighbors = 238, the UMAP projection provides a clear improvement over PCA-2D for $k \sim 135$ (mean $\overline{SD}'_k < 1$), offering a good trade-off for visualization in only 2 dimensions while being able to maintain the global structure of the data, in particular the 6 molecular groups previously identified. This observation highlights the importance of varying the n_neighbors parameter according to the purpose of the projection. Some analyses would require the local structure of the sample neighborhood to be maintained, while others, the global structure.

Preservation of spatial auto-correlations

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

To evaluate the preservation of the spatial auto-correlations, we ranked the top N genes based on the mean MI values for these 3 cases and calculated the overlap between the lists (Fig. 4B). We found that the PCA-5D is only slightly more conservative of the spatial auto-correlations found in the original space than UMAP (unilateral paired t-test $P = 2.2 \times 10^{-16}$). For example, for N = 1,000 (see Euler diagram inserted in Fig. 4B), 88.8% of the genes with the highest MI overlap between the PCA-5D, UMAP, and the original space.

Reuse potential

An interactive TumorMap

Newton and colleagues have recently developed a portal called TumorMap [13, 50], an online tool dedicated to omics data visualization. This new type of integrated genomics portal uses the Google Maps technology designed to facilitate visualization, exploration, and basic statistical interrogation of high dimensional and complex datasets. The pan-LNEN molecular map that we generated in this work (Fig. 3) has been shared on the TumorMap platform. Along with the molecular map, the main clinical, histopathological and molecular features highlighted in the previous studies were uploaded as attributes. The interface enables users to explore and navigate through the map: zooming in and out, coloring and filtering samples based on attributes. The users can also create their own attributes based on pre-existing ones by using operators such as union or intersection. In addition, multiple statistical tests are pre-implemented and available, for example: comparison of attributes without considering the samples positions on the map, comparison of attributes considering samples' positions on the map, and ordering attributes on the basis of their potential to differentiate 2 groups of samples. The interactive nature of the map and the fact that its manipulation does not require computational expertise, could enable the generation of new hypotheses and expand the reuse potential of the dataset.

An interactive computational notebook

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

Integration of new samples

The homogenized read counts of the pan-LNEN data are available on GitHub [14]. Along with the available code, these data could be used to integrate new samples for which RNA-Seq data are available. The raw read counts of the new samples should firstly be generated following the same processing steps described in the section "Data quality controls" (Fig. 1, middle panel) and integrated to the pan-LNEN read counts. We also provide in the Nextjournal notebook, the Nextflow command lines allowing to obtain the read counts. The vst (DESeq2 [39]) should then be applied on the combined dataset and UMAP should finally be rerun to project all samples together in a 2D space. All together, we provide the resources to integrate additional samples into our molecular map, starting from raw sequencing read counts.

Discussion

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

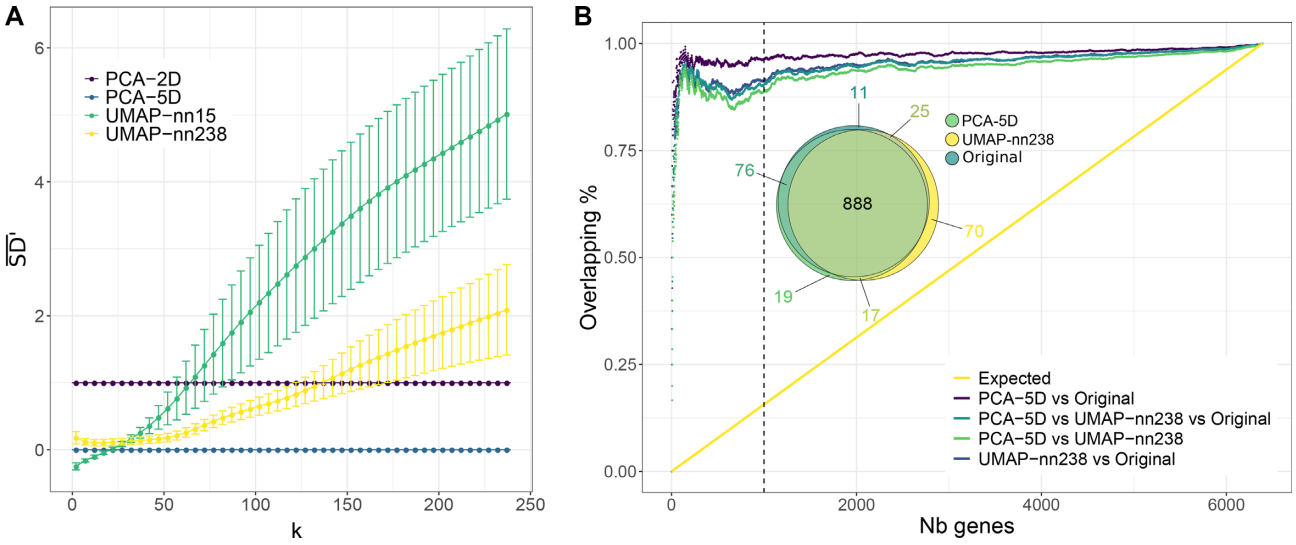


Figure 4: Quality controls performed on the UMAP projection. **(A)** Comparison of the samples' neighborhood preservation for UMAP, PCA-2D, and PCA-5D dimensionality reductions. $\overline{SD}'_k$ values are represented as a function of the number k of nearest neighbors considered, for different dimensionality reduction methods: PCA-2D in purple, PCA-5D in blue, UMAP with n_neighbors = 238 (UMAP-nn-238) in yellow, and UMAP with the default value n_neighbors = 15 (UMAP-nn-15) in green. Error bars correspond to the means ± standard deviations computed across 1,000 replicate simulations. **(B)** Concordance between gene expressions' spatial auto-correlations in the original space, UMAP-nn-238, and PCA-5D dimensionality reductions. For each space, the genes were ranked on the basis of the spatial auto-correlations of their expression (mean MI values). The concordance is measured as the proportion of overlap between the top N genes in the different spaces (colored lines). The yellow line corresponds to the proportion of overlap expected under the null hypothesis (based on the expected mean of the hypergeometric law). The Euler diagram represents the overlaps between the top 1,000 features (N = 1,000, dashed line) resulting from the 3 spaces.

source to improve the molecular characterization of LNEN tumors. Notably, we show here the perfect concordance of the 3 molecular clusters of pulmonary carcinoids independently identified in [7] and [8], validating the discoveries made by these 2 studies and proving the usefulness of this integrative approach.

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

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

Conclusion

Here we provide a molecular map based on homogenized transcriptomic data available for the 4 types of LNENs from 6 different studies. We show that this map represents well both the local and global structure of the data and captures the main biological features previously reported. We provide a full spectrum of data and tools to maximize reuse potential for a wide range of users: raw sequencing reads, gene expression matrix, bioinformatics pipelines, interactive computational notebooks, and an interactive TumorMap. In particular, we indicate how one can update the molecular map by integrating new samples starting from raw sequencing reads. Considering the small sample sizes of molecular studies on rare LNENs, promoting data integration will empower more reliable statistical testing, and this map will therefore serve as a reference in future studies.

Availability of Supporting Data and Materials

R codes used for this article are available in the GigaDB data repository [56]. The data used in this manuscript are available in the European Genome-phenome Archive (EGA), which is hosted at the EBI and the Centre for Genomic Regulation (CRG), under

the accession numbers EGAS00001003699, EGAS00001000650, EGAS00001000925, EGAS00001000708, as well as on Gene expression Omnibus (GEO) under GEO SuperSeries GSE118131.

Ethical Approval

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

Additional Files

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

Abbreviations

AC: atypical carcinoids; ABRA: Assembly-Based Realigner; BAM: Binary Alignment Map; CDS: coding sequence; CGR: Center for Genomic Regulation; CpG: cytosine–phosphate–guanine; CTAT: Trinity Cancer Transcriptome Analysis Toolkit; dbSNP: Single Nucleotide Polymorphism Database; EGA: European Genome-phenome Archive; EMBL-EBI: European Bioinformatics Institute; GATK: Genome Analysis Toolkit; IARC: International Agency for Research on Cancer; LCNEC: large-cell neuroendocrine carcinoma; LCNEC/SCLC-like: large-cell neuroendocrine carcinomas with the molecular features of small cell lung cancers; LNEN: lung neuroendocrine neoplasm; Mb: megabase pairs; MI: Moran index; NEC: neuroendocrine carcinoma; NEN: neuroendocrine neoplasm; NET: neuroendocrine tumor; PCA: principal component analysis; QC: quality control; RNA-Seq: RNA sequencing; SCLC: small-cell lung cancer; SCLC/LCNEC-like: small-cell lung cancers with the molecular features of large-cell neuroendocrine carcinomas; SCLC/SCLC-like: small-cell lung cancers with the molecular features of small-cell lung cancers; SD: Sequence Difference view metric; SNP: single-nucleotide polymorphism; STAR: Spliced Transcripts Alignment to a Reference; TC: typical carcinoids; TES: transcription end site; TSS: transcription start site; UCSC: University of California Santa Cruz; UMAP: Uniform Manifold Approximation and Projection; vst: variance-stabilized transformation; WES: whole-exome sequencing; WGS: whole-genome sequencing.

Competing Interests

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

Funding

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

Authors' Contributions

M.F. and L.F.C. conceived and designed the study. A.A.G.G., E.M., N.A., L.M., and C.V. performed the analyses. V.C. and A.G. gave scientific input for the methylation part. J.D.M. helped with logistics and gave scientific input. A.A.G.G., E.M., N.A., M.F., and L.F.C. wrote the manuscript. All the authors read and commented the manuscript.

Acknowledgments

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

References

1. Rindi G, Klimstra DS, Abedi-Ardekani B, 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. Mod Pathol 2018;**31**(12):1770–86.
2. Peifer M, Fernández-Cuesta L, Sos ML, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet 2012;**44**(10):1104–10.
3. Rudin CM, Durinck S, Stawiski EW, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet 2012;**44**(10):1111–6.
4. Fernandez-Cuesta L, Peifer M, Lu X, et al. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. Nat Commun 2014;**5**:3518.
5. George J, Lim JS, Jang SJ, et al. Comprehensive genomic profiles of small cell lung cancer. Nature 2015;**524**(7563):47–53.
6. George J, Walter V, Peifer M, et al. Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. Nat Commun 2018;**9**(1):1048.
7. Alcala N, Leblay N, Gabriel AAG, et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. Nat Commun 2019;**10**(1):3407.
8. Laddha SV, Da Silva EM, Robzyk K, et al. Integrative genomic characterization identifies molecular subtypes of lung carcinoids. Cancer Res 2019;**79**(17):4339–47.
9. Pelosi G, Bianchi F, Dama E, et al. Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. Virchows Archiv 2018;**472**(4):567–77.
10. Rekhtman N, Pietanza MC, Hellmann MD, et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma–like and

non–small cell carcinoma–like subsets. Clin Cancer Res 2016;**22**(14):3618–29.
11. Simbolo M, Barbi S, Fassan M, et al. Gene expression profiling of lung atypical carcinoids and large cell neuroendocrine carcinomas identifies three transcriptomic subtypes with specific genomic alterations. J Thorac Oncol 2019;**14**(9):1651–61.
12. Fernandez-Cuesta L, Foll M. Molecular studies of lung neuroendocrine neoplasms uncover new concepts and entities. Transl Lung Cancer Res 2019;**8**(S4):S430–4.
13. Newton Y, Novak AM, Swatloski T, et al. TumorMap: Exploring the molecular similarities of cancer samples in an interactive portal. Cancer Res 2017;**77**(21):e111–4.
14. IARCbioinfo/DRMetrics GitHub repository. https://github.com/IARCbioinfo/DRMetrics. Accessed January 2020.
15. Tommaso PD, Floden EW, Magis C, et al. Nextflow, an efficient tool to improve computation numerical stability in genomic analysis. Biol Aujourdhui 2017;**211**(3):233–7.
16. IARCbioinfo/alignment-nf GitHub repository. https://github.com/IARCbioinfo/alignment-nf. Accessed March 2018.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**(14):1754–60.
18. Faust GG, Hall IM. SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics 2014;**30**(17):2503–5.
19. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: Fast processing of NGS alignment formats. Bioinformatics 2015;**31**(12):2032–4.
20. Mose LE, Wilkerson MD, Hayes DN, et al. ABRA: Improved coding indel detection via assembly-based realignment. Bioinformatics 2014;**30**(19):2813–5.
21. Andrews S, Krueger F, Segonds-Pichon A, et al. FastQC. Babraham. 2012. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed August 2019.
22. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 2016;**32**(2):292–4.
23. IARCbioinfo/fastqc-nf GitHub repository. https://github.com/IARCbioinfo/fastqc-nf. Accessed August 2019.
24. IARCbioinfo/qualimap-nf GitHub repository. https://github.com/IARCbioinfo/qualimap-nf. Accessed August 2019.
25. Ewels P, Magnusson M, Lundin S, et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;**32**(19):3047–8.
26. Krueger F. Trim Galore. 2012. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed March 2018.
27. CTAT libraries. https://data.broadinstitute.org/Trinity/CTAT_RESOURCE_LIB/. Accessed May 2020.
28. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013;**29**(1):15.
29. Mose LE, Perou CM, Parker JS. Improved indel detection in DNA and RNA via realignment with ABRA2. Bioinformatics 2019;**35**(17):2966–73.
30. Depristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;**43**(5):491–501.
31. Van der Auwera GA, Carneiro MO, Hartl C, et al. From fastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;**43**(1110):11.10.1–33.
32. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;**33**(3):290–5.
33. Wang L, Wang S, Li W. RSeQC: Quality control of RNA-seq experiments. Bioinformatics 2012;**28**(16):2184–5.
34. Anders S, Pyl PT, Huber W. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics 2015;**31**(2):166.
35. IARCbioinfo/RNAseq-nf GitHub repository. https://github.com/IARCbioinfo/RNAseq-nf. Accessed May 2020.
36. IARCbioinfo/abra-nf GitHub repository. https://github.com/IARCbioinfo/abra-nf. Accessed May 2020.
37. IARCbioinfo/BQSR-nf GitHub repository. https://github.com/IARCbioinfo/BQSR-nf. Accessed May 2020.
38. IARCbioinfo/RNAseq-transcript-nf GitHub repository. https://github.com/IARCbioinfo/RNAseq-transcript-nf. Accessed May 2020.
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;**15**(12):550.
40. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;**30**(10):1363–9.
41. IARCbioinfo/Methylation_analysis_scripts GitHub repository. https://github.com/IARCbioinfo/Methylation_analysis_scripts. Accessed July 2019.
42. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol 2016;**17**(1):208.
43. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv 2018:1802.03426.
44. Konopka T. umap: Uniform Manifold Approximation and Projection. 2019. https://CRAN.R-project.org/package=umap.R package version 0.2.4.0. Accessed May 2020.
45. pan-LNEN TumorMap. https://tumormap.ucsc.edu/?p=RCG_lungNENomics/LNEN. Accessed July 2019.
46. Dray S, Dufour AB. The ade4 Package: Implementing the duality diagram for ecologists. J Stat Softw 2007;**22**(4):1–20.
47. Martins RM, Minghim R, Telea AC. Explaining neighborhood preservation for multidimensional projections. In: Borgo R, Turkay C , eds. Computer Graphics and Visual Computing (CGVC). Eurographics Association; 2015, 10.2312/cgvc.20151234.
48. Moran PA. Notes on continuous stochastic phenomena. Biometrika 1950;**37**(1-2):17–23.
49. Paradis E, Schliep K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2018;**35**:526–8.
50. TumorMap site. https://tumormap.ucsc.edu. Accessed January 2020.
51. IARC bioinformatics platform. https://github.com/IARCbioinfo. Accessed January 2020.
52. Nextjournal notebook: A molecular map of lung neuroendocrine neoplasms. https://nextjournal.com/rarecancersgenomics/a-molecular-map-of-lung-neuroendocrine-neoplasms/. Accessed January 2020.
53. Learned K, Durbin A, Currie R, et al. Barriers to accessing public cancer genomic data. Sci Data 2019;**6**(1):98.
54. Stark Z, Dolman L, Manolio TA, et al. Integrating genomics into healthcare: A global responsibility. Am J Hum Genet 2019;**104**(1):13–20.

55. Espadoto M, Hirata NST, Telea AC. Deep learning multidimensional projections. Inf Vis 2020;**19**(3): 247–69.

56. Gabriel AAG, Mathian E, Mangiante L. Supporting data for "A molecular map of lung neuroendocrine neoplasms." GigaScience Database 2020. http://dx.doi.org/10.5524/100781.