



OPEN

Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of Notch, Hedgehog, Wnt, and ErbB signalling

Dorota Anusewicz^{1,2}✉, Magdalena Orzechowska^{1,2} & Andrzej K. Bednarek¹

Lung malignancies comprise lethal and aggressive tumours that remain the leading cancer-related death cause worldwide. Regarding histological classification, lung squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD) account for the majority of cases. Surgical resection and various combinations of chemo- and radiation therapies are the golden standards in the treatment of lung cancers, although the five-year survival rate remains very poor. Notch, Hedgehog, Wnt and ErbB signalling are evolutionarily conserved pathways regulating pivotal cellular processes such as differentiation, proliferation, and angiogenesis during embryogenesis and post-natal life. However, to date, there is no study comprehensively revealing signalling networks of these four pathways in LUSC and LUAD. Therefore, the aim of the present study was the investigation profiles of downstream target genes of pathways that differ between LUSC and LUAD biology. Our results showed a few co-expression modules, identified through weighted gene co-expression network analysis (WGCNA), which significantly differentiated downstream signaling of Notch, ErbB, Hedgehog, and Wnt in LUSC and LUAD. Among co-expressed genes essential regulators of the cell cycle, DNA damage response, apoptosis, and proliferation have been found. Most of them were upregulated in LUSC compared to LUAD. In conclusion, identified downstream networks revealed distinct biological mechanisms underlying cancer development and progression in LUSC and LUAD that may diversify the clinical outcome of the disease.

Lung carcinomas remain one of the most aggressive malignancies characterized by the highest mortality rate among men and women worldwide^{1,2}. Regarding histological classification lung squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD) account for the majority of lung tumours in non-small cell carcinomas (NSCLCs). In general, the NSCLCs are treated with surgery, which remains the key treatment option, accompanied by various modalities of chemotherapy and radiation. Nevertheless, the five – year survival rate is very poor³ and patients experience early events of relapse, metastasis and death⁴.

Despite that both LUAD and LUSC, belong to the family of NSCLCs, they seem very distinct in terms of prognosis as well as the composition of gene expression and signalling pathways profiles. Importantly, more and more often they are being considered as separate clinical entities⁵. LUAD comprises about 40% of all lung cancer cases. In the majority of patients, LUAD mostly affects non-smokers but is also observed among smokers. Usually, the tumour is located more peripherally and grows slower than the other types, although it tends to form metastasis at the early stages of the disease. LUSC, in turn, is the second most common lung malignancy among tobacco smokers. Its pathogenesis is strongly associated with airway lesions that arise with smoking and is mostly located in the central parts of the lung. LUSC is also regarded as a very heterogeneous entity, among which two major subtypes may be distinguished such as basaloid and non-basaloid tumours⁶. Interestingly, lung cancer shows one of the most diverse genetic landscape harbouring numbers of mutations and copy number

¹Department of Molecular Carcinogenesis, Medical University of Lodz, 90-752 Lodz, Poland. ²These authors contributed equally: Dorota Anusewicz and Magdalena Orzechowska. ✉email: dorota.jedroszka@umed.lodz.pl

alterations. By its nature, LUAD bears numerous rearrangements referring to tyrosine kinase receptors (RTKs) such as *ALK*, *ROS1* and *RET*, and mutations that affect known oncogenes (*KRAS*, *EGFR*), which in contrast are very rare or absent in LUSC. These aberrancies tend to affect corresponding signalling pathways and cause global deregulation as they are closely interconnected by cross-talk of their members⁷.

Among all signalling pathways there are few major developmental mechanisms such as Notch, Wnt, Hedgehog (Hh), and ErbB that are mainly indicated in the cancer models of Hanahan & Weinberg^{8,9} as well as Vogelstein et al.¹⁰ as superior drivers of the carcinogenesis. Additionally, they have been shown strongly involved in lung organogenesis. Notch signalling is evolutionarily conserved pathway determining cell fate during embryogenesis and postnatal life that regulates many cellular processes such as proliferation, differentiation and epithelial-to-mesenchymal transition (EMT), which deregulation may promote carcinogenesis¹¹. Specifically for lungs, Notch determines the fate of proximodistal cells at early stages of organ development. Besides, Notch signalling directs later cytodifferentiation of stem and progenitor cells of specific lineages localized within different segments of airways. Notch is also regarded as a key regulator of cellular differentiation in the parenchyma and vascular compartments, thus it may coordinate alveolar epithelium development and capillary formation. Considering the undeniable role of Notch in lung development, as well as given its significance in supervision over cellular proliferation, differentiation and apoptosis, the involvement of this pathway in lung response to injury becomes apparent. Such tight regulation of Notch activation in various lung cell types has also detrimental reflection found in many pathological states such as chronic obstructive pulmonary disease (COPD), pulmonary fibrosis and indeed, lung cancer that arises from inappropriate Notch signalling¹². Process of lung organogenesis covers likewise epithelial-mesenchymal interactions including cell-cell and cell-matrix interactions necessary for appropriate branching morphogenesis that are in major part dependent on Wnt/ β -catenin signalling. Thus, imbalanced location and duration of β -catenin signalling may significantly affect differentiation of the epithelium and the mesenchyme^{13,14}. Hh is another developmental pathway appearing to maintain stem cells and response to injuries during adulthood. In lungs, its ligands were found during the formation of the tracheobronchial tree. In addition, Hh participates in the regulatory loop through cooperation with Notch and Wnt promoting differentiation of airway epithelial progenitors to form neuro- and non-neuroendocrine lung components¹⁵. ErbB pathway, often identified as EGFR-associated pathway transduces signals through the family of four RTKs comprising *Her1/ErbB-1* (*EGFR*, first discovered member) and *Her2-Her4/ ErbB-2-4*, and was identified as a key regulator of lung maturation as well as maintenance of physiological respiratory functions. In humans, lung alveoli are lined by two types of pneumocytes: type I (95% of the alveolar surface) and type II. *EGFR* and its corresponding ligand *EGF* are expressed in type II pneumocytes that secrete pulmonary surfactant. Activation of *EGFR* was likewise specified to modulate the expression of *MUC5A*, the major component of airway mucus, in response to reactive oxygen species (ROS). Moreover, mucin secretion itself increases with *EGFR* activation. Other studies showed that delivery of anti-EGF antisense oligodeoxynucleotides in vivo caused numerous defects in type II epithelial cells and reduced branching morphogenesis in embryos, hence highlighting the important role of proper ErbB signalling in developing lungs. Remaining members of ErbB family with their ligands (*ErbB-2-4* and *TGF- α* , *HB-EGF*, epiregulin and neuregulins) may be found in type I pneumocytes and therein involved in differentiation and developmental processes as well as damage repair through promoting cellular growth¹⁶. Regarding the prominent role of the pathways discussed in many developmental processes, moreover, in a lung-specific manner, any deregulation during adult homeostasis can lead to various events ultimately leading to the formation of neoplasia. Noteworthy, to date, no study comprehensively describing Notch, Wnt, Hh and ErbB in lung carcinomas in terms of functional cross-talk between their downstream effectors. Moreover, none of them has been correlated with lung cancer subtype, especially at the transcriptome level. Both LUSC and LUAD are members of NSCLC group, although recently they are more often being considered as distinct clinical entities. Therefore, our study aimed to differentiate LUSC and LUAD subtypes at the molecular level and thus reveal functional networks arising from aberrant Notch, Wnt, Hh and ErbB signalling accompanied by correlation with clinical characteristics.

Results

Preliminary analysis on pathway alterations in lung cancer—gene set enrichment analysis. Gene Set Enrichment Analysis (GSEA) was performed as preliminary analysis to examine the differentiation in superior signalling pathways between two subtypes of lung carcinoma – LUAD and LUSC. All these highly conserved signal transduction pathways involved in development and tissue homeostasis were enriched in LUSC vs LUAD: PID_NOTCH_PATHWAY (FDR=0.153, Normalize Enrichment Score (NES)=1.60), KEGG_WNT_SIGNALING_PATHWAY (FDR=0.173, NES=1.56), KEGG_HEDGEHOG_SIGNALING_PATHWAY (FDR=0.101, NES=1.69) and KEGG_ERBB_SIGNALING_PATHWAY (FDR=0.180, NES=1.54) (Fig. 1).

Downstream effects of aberrant signalling in LUSC and LUAD through Notch, Hh, ErbB and Wnt. According to GSEA results, we decided to focus on downstream effects of signalling through Notch, Hh, ErbB and Wnt pathways, therefore subsequent analyses considered alterations in target genes of pathway-specific transcription factors, separately for each of the pathways.

Alterations in signalling networks associated with Notch, Hh, ErbB and Wnt downstream effects in LUSC and LUAD—Weighted Gene Co-expression Network Analysis (WGCNA). We investigated major biological differences between LUSC and LUAD in signalling networks downstream to Notch, Hh, ErbB and Wnt via WGCNA to find modules of genes that shared common expression profiles among pathway-specific transcription factor's targets. We analysed The Cancer Genome Atlas (TCGA) data of gene expression of 499 LUSC and 515 LUAD patients. Based on WGCNA pipeline, we identified a total of 9, 12, 9 and

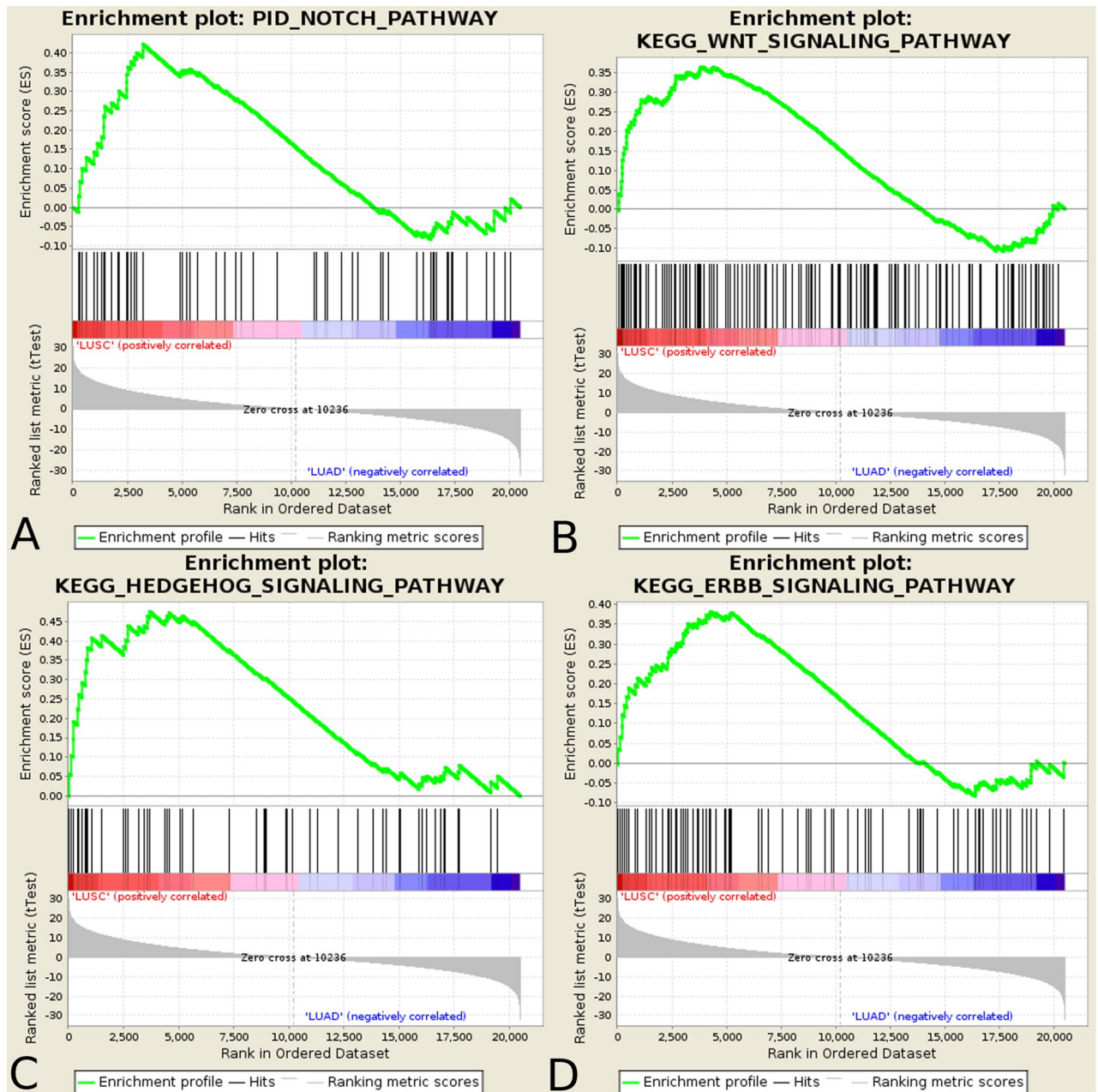


Figure 1. Signalling pathways enriched in lung squamous cell carcinoma and lung adenocarcinoma. Enrichment plots present (A) Notch pathway, (B) Wnt pathway, (C) Hh pathway, and (D) ErbB pathway.

11 distinct co-expression modules for Notch, Hh, Wnt and ErbB pathway, respectively. In all cases, uncorrelated genes were assigned to the grey module and were excluded from the subsequent investigations.

To understand the biological characteristic of genes from the modules, we performed the correlation analysis between module eigengene (ME) considered as the most representative gene expression profile (the first principle component) of the module and LUSC/LUAD phenotype. According to the module-trait relationship analysis in targets of Notch pathway (Fig. 2A), genes clustered in brown ($r=0.77$, $p\text{-value}=3e-200$), turquoise ($r=0.62$, $p\text{-value}=4e-110$) and yellow ($r=0.53$, $p\text{-value}=6e-73$) modules showed the strongest positive correlation with phenotype. Regarding Hh targets 4 modules were significantly associated with LUSC/LUAD type (Fig. 3A): brown ($r=0.81$, $p\text{-value}=2e-235$), blue ($r=0.67$, $p\text{-value}=3e-132$), red ($r=0.54$, $p\text{-value}=5e-78$) and purple ($r=0.5$, $p\text{-value}=5e-65$). Purple and turquoise modules ($r=0.51$, $p=1e-67$; $r=0.73$, $p=5e-172$, respectively) were of the strongest positive correlation with LUSC/LUAD phenotype among ErbB targets (Fig. 4A) and blue and green modules ($r=0.76$, $p=1e-187$; $r=0.59$, $p=5e-98$, respectively) among Wnt pathway effectors (Fig. 5A). These MEs were also the most promising as exhibiting the highest gene significance across all modules (Figs. 2, 3, 4, 5, B section).

For each gene in a module, we performed an analysis of the correlation between module membership (MM) and Gene Significance (GS), separately for each module. MM, which is defined as a correlation of gene expression

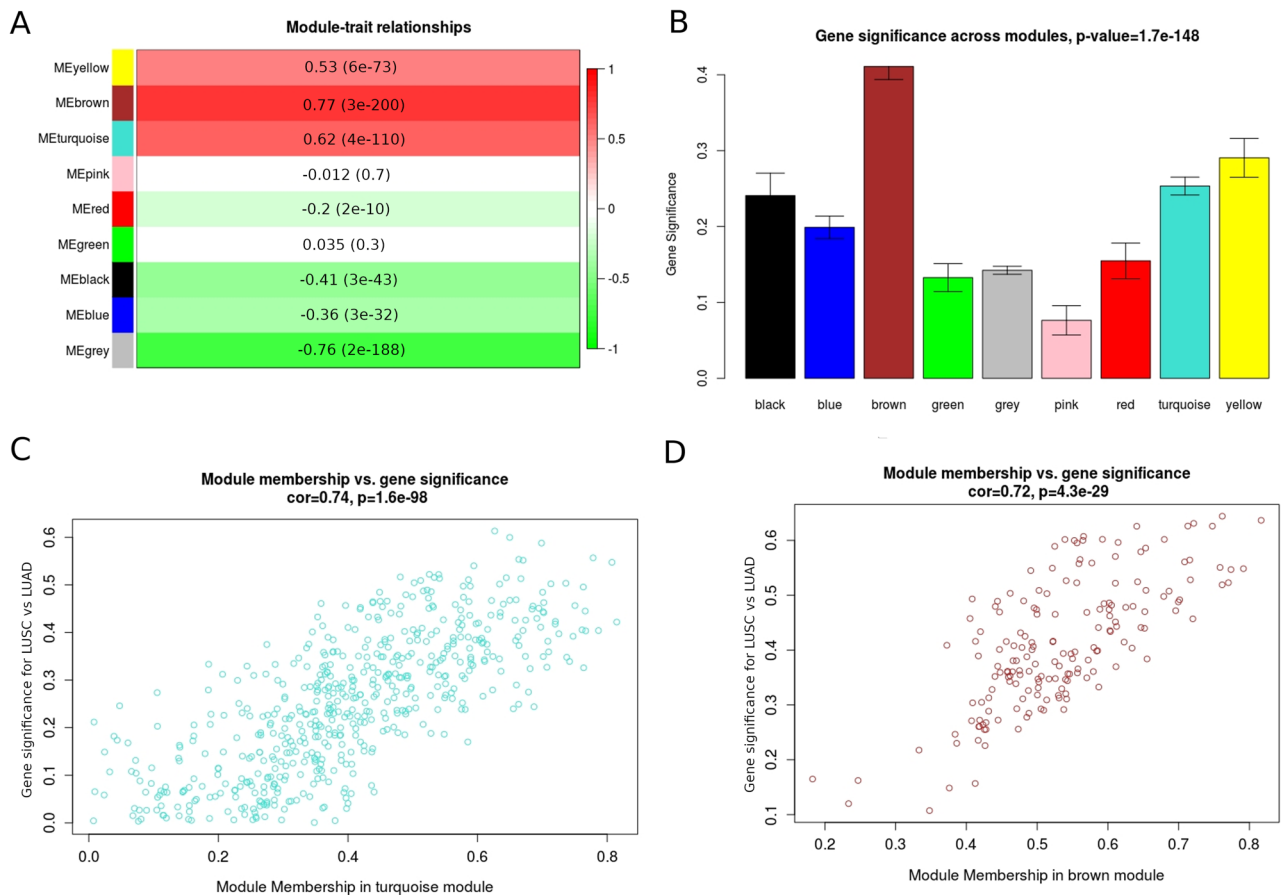


Figure 2. Differential expression of Notch downstream targets vastly characterizing lung cancer subtypes within turquoise and brown modules. These groups were mostly correlated with trait of interest (A,B) and showed the highest importance of genes within the particular module (C,D).

profile with ME (the first principal component of a module), showed the importance of a particular gene across the module. As demonstrated in the scatter plots, among Notch pathway targets MM were highly associated with GS in brown and turquoise modules (brown: $\text{cor}=0.72$, $p=4.3\text{e-}29$; turquoise: $\text{cor}=0.74$, $p=1.6\text{e-}98$) (Fig. 2C and 2D) and the significant correlation was reported regarding Hh pathway targets for brown ($\text{cor}=0.82$, $p=2.7\text{e-}52$), blue ($\text{cor}=0.73$, $p=2.6\text{e-}61$), purple ($\text{cor}=0.73$, $p=2.9\text{e-}07$) and red ($\text{cor}=0.54$, $p=1.4\text{e-}08$) modules (Fig. 3C–F). We have also found significant associations between MM and GS in blue module in Wnt pathway targets ($\text{cor}=0.74$, $p=1.3\text{e-}86$; Fig. 5C) as well as purple ($\text{cor}=0.83$, $p=1.2\text{e-}10$) and turquoise ($\text{cor}=0.71$, $p=1\text{e-}126$) modules across ErbB pathway effectors (Fig. 4C, D). This analysis highlighted the essential character of particular elements among the modules that tended to show strong association with lung cancer subtype, i.e. LUSC and LUAD.

Finally, heatmaps generated for modules of high relevance (brown, turquoise, yellow for Notch targets, brown, blue, purple, red for Hh targets, blue and green for Wnt targets, purple and turquoise for ErbB targets) reflected differential biology of LUAD and LUSC regarding the expression of genes therein (Fig. 6 and Supplementary Figs. 1–4).

Functional annotation and enrichment analysis of WGCNA modules in Notch, Hedgehog, Wnt and ErbB pathway downstream targets.

We summarized the findings of WGCNA with functional annotation of the modules to provide an understanding of biological mechanisms associated with the genes clustered in modules: blue, green of Wnt pathway targets, turquoise, purple of ErbB pathway targets, brown, turquoise, yellow of Notch pathway targets and brown, blue, purple and red of Hh pathway targets. We chose two main categories—KEGG canonical pathways and GO biological processes, both derived from MSigDB. Majority of genes across modules were associated with proliferative and repair processes such as cell cycle, cytoskeleton organization and biogenesis, base excision repair, mismatch repair, MAPK signalling, response to stress as well as developmental processes such as intracellular transport and multicellular organismal development. A detailed description of major findings is shown in Supplementary Table 1.

Dimensional characteristics of LUSC and LUAD patients—Multiple Factor Analysis (MFA). To confirm the relationship between the groups of variables (expression of downstream effectors chosen from the

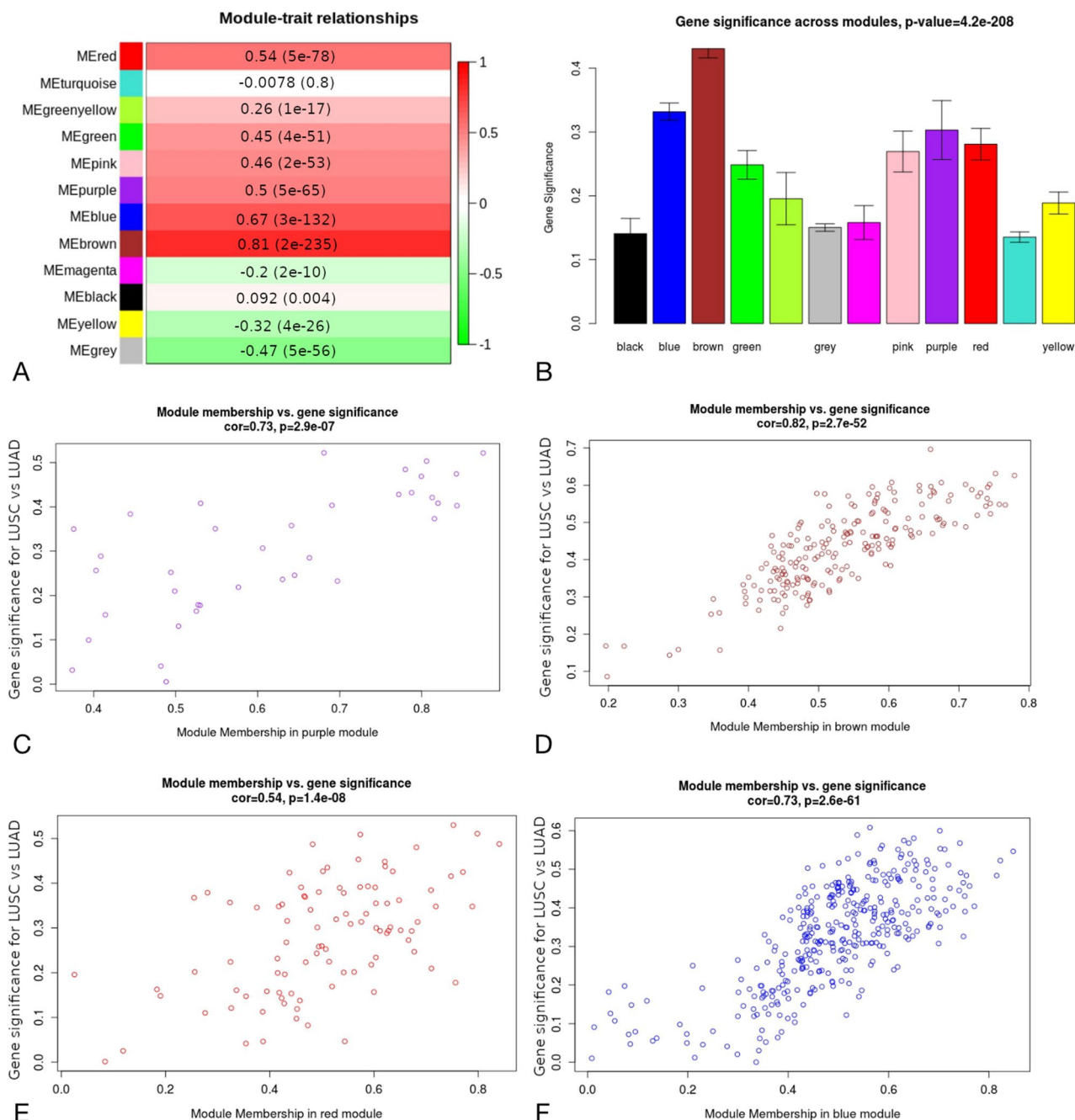


Figure 3. Differential expression of Hh downstream targets vastly characterizing lung cancer subtypes within purple, brown, red and blue modules. These groups were mostly correlated with trait of interest (A,B) and showed the highest importance of genes within the particular module (C–F).

highly significant modules and lung cancer subtypes) describing the individuals (patients) concerning their clinical characteristics, we applied Multiple Factor Analysis (MFA). As expected, we found significant partitioning of patients across the first dimension for LUSC/LUAD phenotype with 19.24% variance among Notch effectors, 24.02% variance among Hh effectors, 22.12% variance among Wnt effectors and 28.1% variance among ErbB effectors (Fig. 7).

Association of the gene expression with patients survival. Additionally, to identify genes that may be of diagnostic-prognostic or therapeutic importance, we analyzed the relationship between the most interesting target genes among the modules with clinical outcomes such as overall survival (OS) and disease-free survival (DFS). Kaplan–Meier survival analysis of the groups of LUSC and LUAD patients was performed using algorithm determining the optimal cutpoint splitting patients into subgroups of more/less favorable outcomes. Of the results, we focused on the genes that proved to have a significant but distinct impact on OS either DFS in

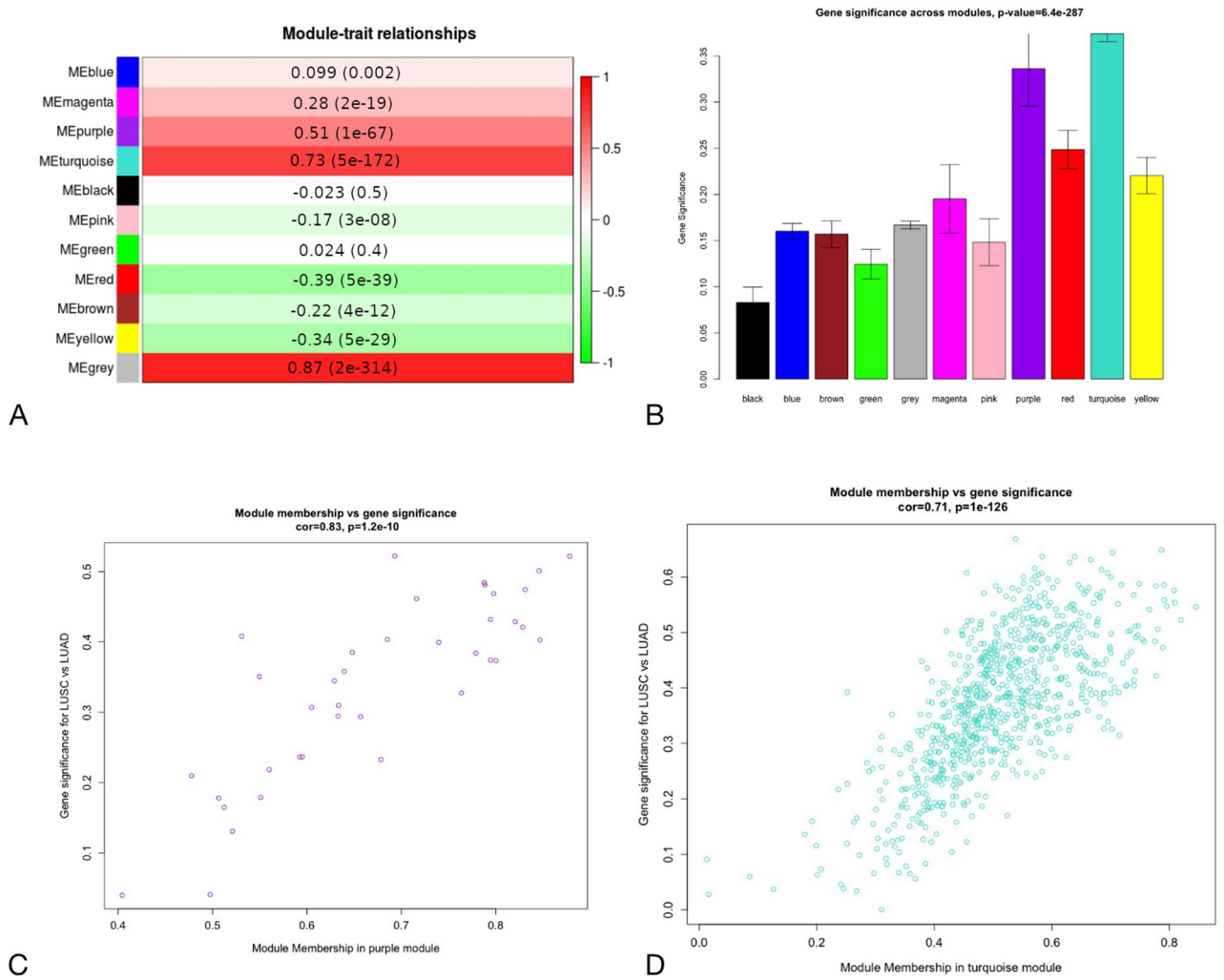


Figure 4. Differential expression of ErbB downstream targets vastly characterizing lung cancer subtypes within purple and brown modules. These groups were mostly correlated with trait of interest (A,B) and showed the highest importance of genes within the particular module (C,D).

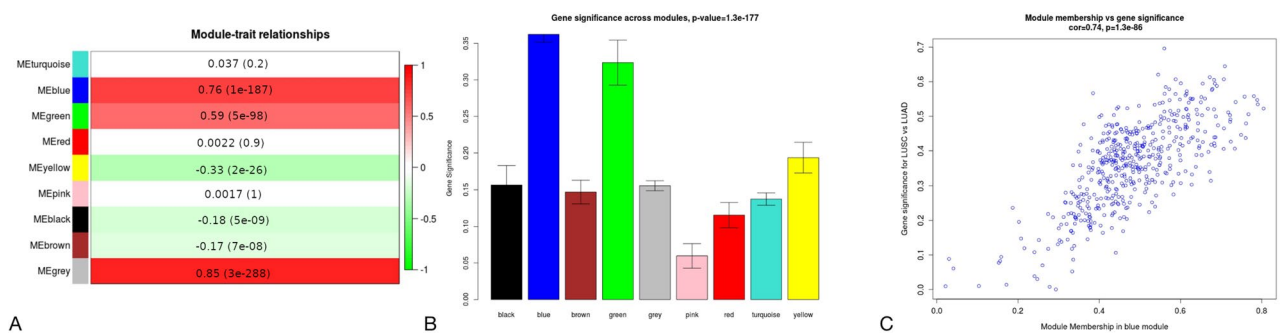


Figure 5. Differential expression of Wnt downstream targets vastly characterizing lung cancer subtypes within blue module. This group was mostly correlated with trait of interest (A,B) and showed the highest importance of genes within the blue module (C).

both LUAD and LUSC. Regarding OS, we identified genes such as *CDC25A*, *CDK2*, *E2F8*, *KIF11*, *KIF2A*, *KIF4A*, *MAPK8*, *MCM5*, *MCM6*, *MYC*, *PARP1*, and *PIK3CA* that significantly differed in the patients' outcome between LUSC and LUAD. Additionally, we identified unique genes significantly altering the prognosis only in one of the tumors such as *CDK16* and *MSH2* for LUSC and *BRCA1*, *BRCA2*, *CCNB1*, *CDKN3*, *E2F1*, *E2F2*, *KIF14*, *KIF23*, *MAPK6*, *MCM2*, *MCM4*, *MCM8*, *MCM10*, and *RAD51* for LUAD. Regarding DFS, we identified the unique genes such as *BRCA1*, *BRCA2*, *KIF14*, *MCM5*, *MCM8*, *MSH2*, *PARP1*, and *TP63* for LUSC and *E2F2*, *MAPK6*, *MAPK8*, *MCM2* for LUAD (statistics data are shown in Tables 1 and 2).

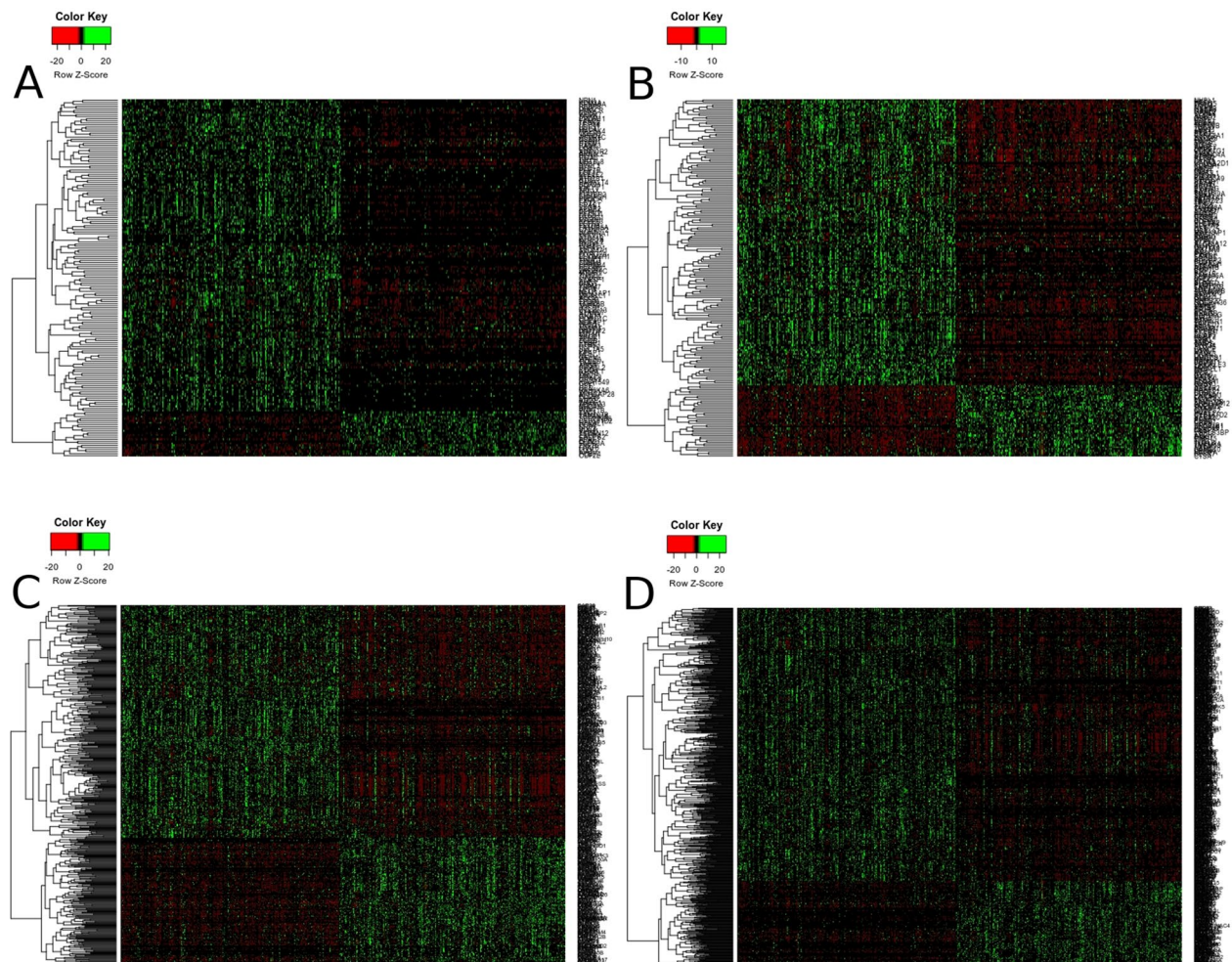


Figure 6. Molecular profiles of (A) Notch (brown module), (B) Hh (brown module), (C) Wnt (blue module), (D) ErbB (turquoise module) pathway downstream targets that differentiate lung squamous cell carcinoma from lung adenocarcinoma.

Discussion

Lung cancer is a very complex and heterogeneous disease, categorized into two major types, small cell lung carcinoma (SCLC) and NSCLC. NSCLCs account for approx. 85% of all lung cancers and is associated with high rates of proliferation and metastases as well as poor prognosis for advanced-stage disease¹⁷. In this study, we focused on LUAD and LUSC among all NSCLCs. LUSC manifests in poorer prognosis and usually arises as a tumour localised in the proximal part of the bronchial tree and is strongly associated with tobacco smoking, whereas LUAD is usually peripherally located and occurs more often among non-smokers¹⁸. Since both subtypes vary in clinical and histopathological features, their molecular mechanism of carcinogenesis and tumour progression may be diverse.

The development of NGS technology has enabled to study the expression level of many genes simultaneously and understand how the networks and pathways interact with each other. To identify the distinct molecular profiles based on gene expression analysis of LUSC and LUAD, we focused on four evolutionary conserved signalling pathways: Notch, Hh, Wnt and ErbB, which tightly regulate proliferation, differentiation, apoptosis, migration and motility. Previous studies showed that the EGFR/ErbB pathway is involved in LUAD pathology with mutations of EGFR and ERBB2 identified in 11% and 3% of LUAD cases, respectively¹⁹. Another group reported differential and clinically important subtypes of LUSC based on gene expression profiles²⁰. Finally, a recent comparison of LUSC and LUAD showed distinct expression profiles of genes involved in tumour immune response^{21,22}. Alternations of Notch, Hh, Wnt and ErbB pathways were repeatedly associated with development and progression of many malignancies. Our GSEA analysis revealed that Notch, Hh, Wnt, and ErbB pathways were significantly overrepresented in LUSC compared to LUAD. Regarding the differential expression of the genes directly involved in these pathways, it seems logical to assume that there are alternations in their downstream target genes, which may influence the molecular and clinical character of LUAD and LUSC. Thereby, identification of their differential signalling in NSCLC may help to understand the biology of lung cancer and lead to new personalized therapies in the future.

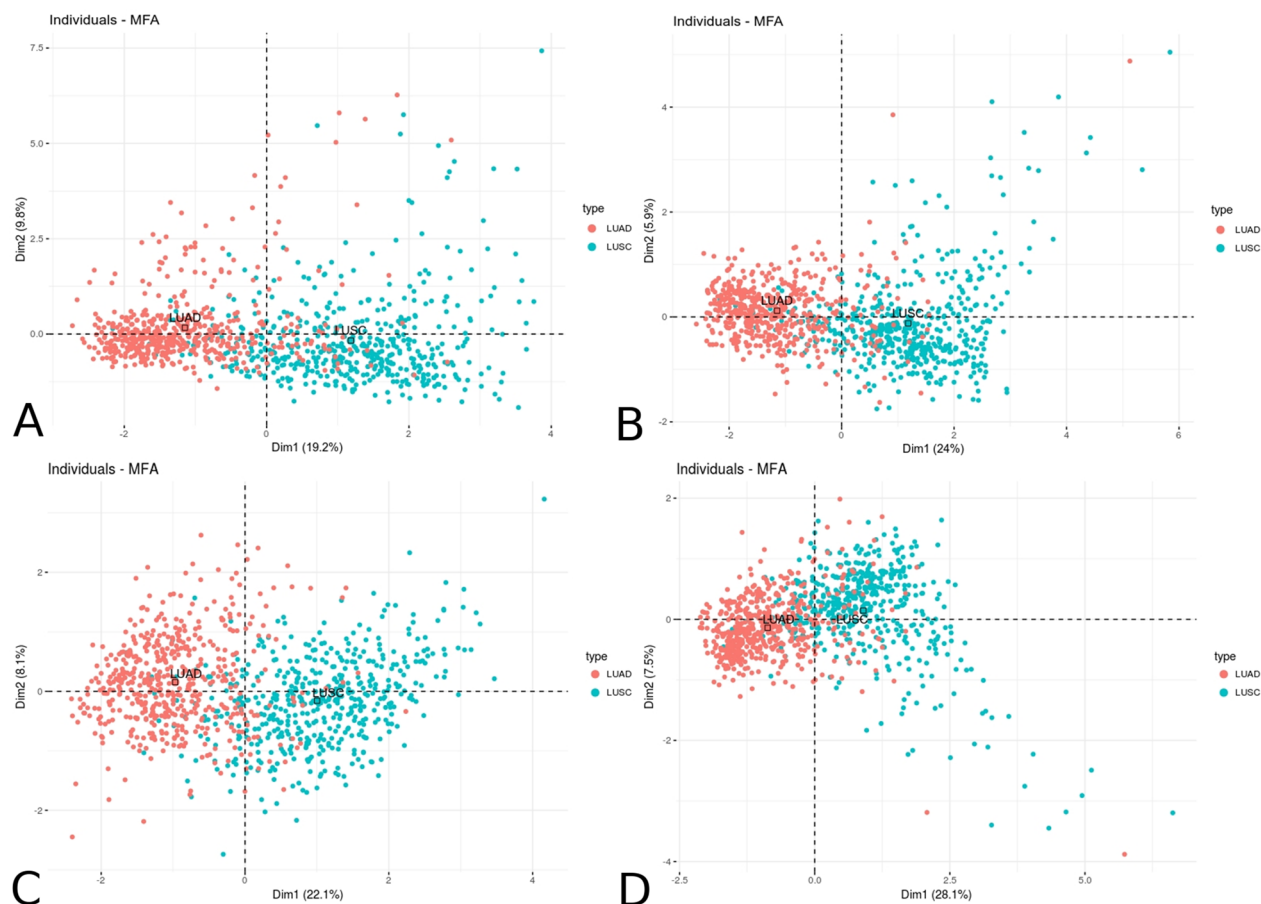


Figure 7. Dimensional partitioning of lung squamous cell carcinoma and lung adenocarcinoma patients according to the resultant expression of the most significant WGCNA modules of (A) Notch, (B) Hh, (C) Wnt and (D) ErbB pathway effectors.

By using WGCNA we compared gene expression profiles between subtypes of lung cancer and revealed 11 co-expression modules of Notch, Wnt, Hh and ErbB downstream targets showing a strong positive correlation ($\text{cor} > 0.5$) with either LUSC or LUAD subtype. These modules corresponded to a total of 1965 differentially expressed genes of potential biological relevance associated with gene ontology terms such as cell cycle, cellular differentiation and proliferation, DNA repair, metabolic processes and apoptosis (Supplementary Table 1) that point out the major biological differences between LUSC and LUAD.

Major differences included genes involved in the cell cycle control and some of them have been already known to be involved in tumorigenesis^{23,24}. Loss of checkpoint and integrity induced by failures of the cell-cycle machinery, triggering uncontrolled tumour proliferation and leading to malignant transformation. In this context, our analysis has shown that many of the cell cycle genes were elevated in LUSC compared to LUAD. Among them we identified cyclins, cyclin-dependent kinases (CDKs), E2F family transcription factors, kinesin superfamily proteins (KIFs) and minichromosome maintenance proteins (MCMs) (Supplementary Figs. 4; Figs. 3 A, C and D), thus revealing different effects of abolished cell cycle control through Notch, Hh, Wnt and ErbB pathways.

The balance of cellular proliferation and apoptosis plays a pivotal role in the control of tumour growth, which as we found, seems to be driven by the effects of abrogated Notch signalling (Supplementary Table 1). Deregulation of apoptosis is implicated in tumour initiation, progression and drug resistance in many human cancers and is also identified as one of the hallmarks of cancer^{8,9}.

Another interesting group of genes are DNA repair genes, which play a key role in genome surveillance and protection. Functional deregulation of DNA repair is one of the common features of highly aggressive human malignancies²⁵. There are at least eight distinct DNA repair mechanisms in human cells, amongst which are: base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), non-homologous end-joining (NHEJ), homologous recombination (HR) and Fanconi anaemia pathway (FA)^{25–27}. Additionally, *BRCA1* and *BRCA2* expression was shown to differentiate LUSC and LUAD, and their reduced expression was previously reported as associated with hypermethylation of gene promoters in LUAD²⁸.

Our analysis has shown that many of the cell cycle genes, including cyclins and CDKs, were elevated in LUSC compared to LUAD. Especially, overexpression of *CDK2* and *CDK16* was shown to cause abnormal regulation of cell cycle and to promote cell proliferation²⁹. Our findings are in compliance with previous research that indicated higher expression of *CDKN3* in LUSC than in LUAD^{30,31}, whereas overexpression of cyclin B1 (*CCNB1*, a key

| GENE | LUSC | | | | | LUAD | | | | |
|---------------|--------|---------|----------|-----------|-----------|------|----------|----------|-----------|-----------|
| | HR | P-value | Cutpoint | <Cutpoint | >Cutpoint | HR | P-value | Cutpoint | <Cutpoint | >Cutpoint |
| <i>CDC25A</i> | 0.608 | 0.00388 | 295.5 | 317 (64%) | 182 (36%) | 1.71 | 0.00322 | 131.9 | 365 (71%) | 150 (29%) |
| <i>CDK2</i> | 0.613 | 0.0302 | 1530 | 425 (85%) | 74 (15%) | 1.54 | 0.0229 | 617.8 | 243 (47%) | 272 (53%) |
| <i>E2F8</i> | 2.11 | 0.00841 | 71.46 | 61 (12%) | 438 (88%) | 1.54 | 0.0395 | 255.8 | 425 (83%) | 90 (17%) |
| <i>KIF11</i> | 0.652 | 0.0102 | 1116 | 296 (59%) | 203 (41%) | 1.92 | 0.00132 | 353.7 | 201 (39%) | 314 (61%) |
| <i>KIF2A</i> | 1.52 | 0.0234 | 500.7 | 173 (35%) | 326 (65%) | 1.7 | 0.0156 | 375.3 | 104 (20%) | 411 (80%) |
| <i>KIF4A</i> | 0.713 | 0.0467 | 831.2 | 312 (63%) | 187 (37%) | 1.68 | 0.00457 | 456.3 | 336 (65%) | 179 (35%) |
| <i>MAPK8</i> | 1.82 | 0.0114 | 193.3 | 112 (22%) | 387 (65%) | 1.45 | 0.0483 | 215.8 | 236 (46%) | 279 (54%) |
| <i>MCM5</i> | 1.41 | 0.0489 | 1891 | 161 (32%) | 338 (68%) | 1.78 | 0.00128 | 1251 | 319 (62%) | 196 (38%) |
| <i>MCM6</i> | 0.624 | 0.0121 | 2080 | 369 (74%) | 130 (26%) | 1.64 | 0.00864 | 1048 | 244 (47%) | 271 (53%) |
| <i>MYC</i> | 1.45 | 0.0407 | 2308 | 167 (33%) | 332 (67%) | 1.91 | 0.000687 | 2026 | 395 (77%) | 120 (23%) |
| <i>PARP1</i> | 0.0104 | 0.0104 | 4448 | 190 (38%) | 309 (62%) | 2.82 | 0.000176 | 7323 | 482 (94%) | 33 (6%) |
| <i>PIK3CA</i> | 0.0487 | 0.0487 | 649 | 168 (34%) | 331 (66%) | 2.18 | 1.55E-05 | 429.2 | 333 (65%) | 182 (35%) |
| <i>CDK16</i> | 0.56 | 0.011 | 1851 | 52 (10%) | 447 (90%) | | | | | |
| <i>MSH2</i> | 0.626 | 0.0113 | 1206 | 340 (68%) | 159 (32%) | | | | | |
| <i>BRCA1</i> | | | | | | 1.69 | 0.0106 | 199.3 | 184 (36%) | 331 (64%) |
| <i>BRCA2</i> | | | | | | 2.06 | 0.00054 | 200.3 | 432 (84%) | 83 (16%) |
| <i>CCNB1</i> | | | | | | 1.94 | 0.000392 | 680 | 261 (51%) | 254 (49%) |
| <i>CDKN3</i> | | | | | | 1.64 | 0.00641 | 144.6 | 292 (57%) | 223 (43%) |
| <i>E2F1</i> | | | | | | 1.62 | 0.0263 | 268.1 | 176 (34%) | 339 (66%) |
| <i>E2F2</i> | | | | | | 1.65 | 0.0175 | 84.59 | 167 (32%) | 348 (68%) |
| <i>KIF14</i> | | | | | | 1.87 | 0.00051 | 195.5 | 307 (60%) | 208 (40%) |
| <i>KIF23</i> | | | | | | 1.82 | 0.000855 | 433.7 | 334 (65%) | 181 (35%) |
| <i>MAPK6</i> | | | | | | 2.22 | 3.31E-05 | 1594 | 405 (79%) | 110 (21%) |
| <i>MCM10</i> | | | | | | 1.58 | 0.0168 | 108.3 | 219 (43%) | 296 (57%) |
| <i>MCM2</i> | | | | | | 1.52 | 0.0211 | 1016 | 279 (54%) | 236 (46%) |
| <i>MCM4</i> | | | | | | 2.1 | 3.61E-05 | 1413 | 303 (59%) | 212 (41%) |
| <i>MCM8</i> | | | | | | 1.58 | 0.0188 | 177.8 | 177 (34%) | 338 (66%) |

Table 1. Prognostic effect of chosen target genes on OS in lung subtypes.

| GENE | LUSC | | | | | LUAD | | | | |
|--------------|-------|----------|----------|-----------|-----------|-------|---------|----------|-----------|-----------|
| | HR | P-value | Cutpoint | <Cutpoint | >Cutpoint | HR | P-value | Cutpoint | <Cutpoint | >Cutpoint |
| <i>BRCA1</i> | 2.54 | 0.0157 | 435.6 | 201 (40%) | 298 (60%) | | | | | |
| <i>BRCA2</i> | 5.24 | 0.000186 | 386.4 | 465 (93%) | 34 (7%) | | | | | |
| <i>KIF14</i> | 2.54 | 0.0215 | 258.1 | 168 (34%) | 331 (66%) | | | | | |
| <i>MCM5</i> | 3.06 | 0.0264 | 1729 | 117 (23%) | 382 (77%) | | | | | |
| <i>MCM8</i> | 2.86 | 0.000848 | 708.6 | 359 (72%) | 140 (28%) | | | | | |
| <i>MSH2</i> | 2.43 | 0.00718 | 1579 | 420 (84%) | 79 (16%) | | | | | |
| <i>PARP1</i> | 3.48 | 0.000885 | 8435 | 461 (92%) | 38 (8%) | | | | | |
| <i>TP63</i> | 0.382 | 0.0176 | 13,310 | 333 (67%) | 166 (33%) | | | | | |
| <i>E2F2</i> | | | | | | 0.363 | 0.045 | 232.2 | 428 (83%) | 87 (17%) |
| <i>MAPK6</i> | | | | | | 0.522 | 0.033 | 1067 | 262 (51%) | 253 (49%) |
| <i>MAPK8</i> | | | | | | 0.398 | 0.0449 | 315.7 | 406 (79%) | 109 (21%) |
| <i>MCM2</i> | | | | | | 1.58 | 0.0188 | 1268 | 348 (68%) | 167 (32%) |

Table 2. Prognostic effect of chosen target genes on DFS in lung subtypes.

molecule for G2/M phase) was recently identified as a predictive marker of worse overall survival among LUSC patients⁷. Moreover, high expression of *CCNB1* increased cell differentiation, high proliferative index, vascular invasion and thus increased malignant potential^{32,33}. *CDC25A* in turn seemed to possess oncogenic properties and its overexpression was frequently associated with the malignancies and poor prognosis^{34,35}.

Another group of genes controlling the cell cycle comprises the family of E2F transcription factors. One of their abundant functions is a regulation of expression of genes essential for the transition from G1 to S phase of cell cycle like *Cdc25a*, cyclin A and cyclin B³⁶. Our analysis revealed an upregulated expression of *E2F1* and *E2F2* in LUSC compared to LUAD. Several studies reported that E2F factors likely contribute to lung

carcinogenesis³⁷. The overexpression of *E2F1* has been associated with the development of NSCLC and indicated worse prognosis^{38,39}. *E2F2* expression was predominantly elevated in NSCLC tumours and was also shown to correlate with cell proliferation leading to tumour progression³⁷. Park et al. found overexpression of *E2F8* in lung cancer cell lines and in lung cancer tissue samples that was ultimately associated with poorer prognosis. Moreover, in the same study inhibition of *E2F8* suppressed cell proliferation, colony formation and invasion and tumour growth in vitro and in vivo⁴⁰.

Kinesin superfamily is essential for mitosis and meiosis, intracellular transport and cell migration⁴¹. In our analysis, we demonstrated that several genes of the KIF family are heightened in LUSC. To date, it has been established that overexpression of specific genes such as *KIF4A* and *KIF14* is strongly associated with poorer prognosis of NSCLC cases^{42,43}. Significant increase in *KIF2A* expression in NSCLC was additionally associated with lymph node metastasis⁴⁴ as well as LUAD progression⁴⁵. Our study revealed differences in *KIF11*, the upregulation of which was previously reported in LUSC compared to LUAD⁴⁶.

We found that genes of MCM family involved in eukaryotic genome replication were heightened in LUSC vs LUAD (Supplementary Figs. 2–4 and Fig. 6 B, C and D). In the previous reports, MCM genes have been proposed as prognostic biomarkers of proliferation in lung cancer^{47–51}. NSCLCs displayed elevated expression of *MCM2* as well as *MCM5* and *MCM6* that were associated with patients' shorter overall survival. Additionally, the same study reported that higher *MCM5* was significantly correlated with distant metastases⁴⁹. In our research, the expression of all above mentioned MCM genes were higher in LUSC than in LUAD, while only *MCM5* were found to have the same trend in survival. Moreover, higher expression of *MCM4* was associated with non-adenocarcinomas as well as smoking, and we confirmed the above finding demonstrating upregulation of *MCM4* in LUSC⁴⁸.

Cancer cells often show abolished signal transduction that leads to proliferation in response to external signals. Among the most differentiated genes between LUSC and LUAD we found genes involved in frequently deregulated pathways during carcinogenesis, e.g. *MYC*, *TP63*, *GSK-3 β* , *PIK3CA*, *MAP* kinases (Figs. 6 A, B and D and Supplementary Figs. 1, 2 and 4). Amplification of c-myc was shown to play an important role in such processes as metastasis, invasion and resistance to chemotherapy^{52,53}. *TP63* gene amplification and corresponding protein overexpression have been so far documented mainly in squamous cell carcinoma and related to the tumour proliferation fraction^{54,55}. Glycogen synthase kinase-3 (GSK-3) is a key regulator of numerous signalling pathways during embryogenesis and in metabolic control. We found that *GSK-3 β* isoform exhibited higher expression in LUSC patients. Respectively, previous studies have shown that overexpression of the *GSK-3 β* in NSCLC patients regulated cell proliferation, tumorigenesis, apoptosis and cell invasiveness and thus was identified as a risk factor of poor prognosis⁵⁶. *PIK3CA* is one of the most frequently mutated genes in human cancers and its somatic mutations have also been reported in lung cancer⁵⁷. Our analysis showed differential expression of *PIK3CA* in NSCLC but there are no reports of *PIK3CA* gene overexpression being associated with either carcinogenesis or progression of lung cancer. However, elevated expression of *PIK3CA* has been reported in various types of cancer including oesophageal squamous cell carcinoma, colorectal and breast cancer where was related to invasiveness, metastasis and poor prognosis^{58–60}. *MAPK* has been linked to cell proliferation and transformation. In our study diverged expression levels of *MAPK6* and *MAPK8* were found between LUSC and LUAD, with higher expression in LUSC. *MAPK6* (Supplementary Figs. 2–3, Figs. 6 B and C), also known as *ERK3*, were found to be strongly upregulated in human lung carcinoma and promotes cancer cell invasion⁶¹. *MAPK8* (*JNK1*) is involved in transduction of extracellular signals such as growth factors or cytokines and also was to be an important contributor to the tumour promoting activity of tobacco smoke in lungs⁶².

Many of the DNA repair genes were expressed differently in LUSC and LUAD, which may suggest that the development and aggressiveness of these tumours depends on distinct mechanisms. Homologous recombination is an error-free double-strand break (DSB) repair pathway active during the S and G2 phases of the cell cycle due to the necessity of a sister chromatid for use as a homologous template⁶³. Among genes involved in the HR, we identified *RAD51*, *BRCA1* and *BRCA2*. Decreased BRCA expression caused by methylation or mutation has been shown to impair the homologous recombination. In addition, aberrant protein expression as well as low mRNA level of *BRCA1* and *BRCA2*, was significantly associated with promoter hypermethylation of these genes, especially in LUAD patients²⁸. These findings are in compliance with our results demonstrated lowered expression of *BRCA1* and *BRCA2* in LUAD compared to LUSC (Fig. 6D). The above suggests that alterations of the key members of the DSB repair pathway are important primarily in the pathogenesis of LUAD. *RAD51* is a protein, which interacts with various tumour suppressor, including *BRCA1*, *BRCA2*, *TP53*. High level of *RAD51* was related to chemo- and radioresistance of lung cancer^{64,65} as well as the enhanced propensity of cancer cells to survive and avoid apoptosis. Significantly shorter survival was also observed among NSCLC patients with higher *RAD51* expression⁶⁵. In turn, *PARP1* and *PARP2* have been described as active players of the DNA damage response, DNA metabolism and chromatin architecture. They are important in BER recognition of single-stranded breaks. Importantly, our analysis has revealed that both *PARP1* and *PARP2* were upregulated in LUSC compared to LUAD. Chen K. et al. reported that *PARP1* is strongly expressed in metastatic NSCLC and facilitates migration and invasion of NSCLC cells. Moreover, the overall survival was significantly lower in *PARP1* high expression group of patients than in *PARP1* low expression group⁶⁶.

MMR is a highly conserved biological mechanism that recognizes and repairs erroneous insertions, deletions and base substitution that have been neglected by the intrinsic proofreading activity of the DNA polymerases⁶⁷. Inactivation of MMR induces a mutator phenotype and causes a predisposition to cancer. Indeed, loss of *MSH2*, one of the key components of the MMR, influences the enhancement of genomic instability. Our analysis showed reduced expression of *MSH2* in LUAD. Previous studies showed in turn that low expression of *MSH2* was positively correlated with decreased overall survival of lung cancer patients due to increased genome instability, a hallmark of MMR-deficient cells⁶⁸. We found that *MSH2* together with *MSH6* have lower expression in LUAD patients than among LUSC.

Analysis of gene ontology (GOs) and pathways identified diverse gene sets, suggesting that these may serve primarily roles in differentiation of LUSC and LUAD pathogenesis. Moreover, we aimed to evaluate if any of the differentially expressed genes had a potential prognostic impact for LUAD or LUSC. The Evaluate Cutpoint analysis led us to discover a few important genes that could differentially determine disease outcome as well as disease free survival in LUAD versus LUSC. Higher expression of *E2F8*, *MAPK8*, *MCM5* and *MYC* followed the same, negative, prognostic impact on both LUSC and LUAD, however taking into account the number of patients with expression of these genes above cutpoint (in LUSC: 88% for *E2F8*, 78% for *MAPK8*, 68% for *MCM5* and 67% for *MYC*; in LUAD: 17% for *E2F8*, 54% for *MAPK8*, 38% for *MCM5*, 23% for *MYC*), it could be assumed that frequently elevated expression of these genes in patients with LUSC may determine the final outcome. The most interesting from viewpoint of distinguish LUSC from LUAD regarding survival rate are genes with opposite trends such as *CDC25A*, *CDK2*, *KIF11*, *KIF4A*, *MCM6*, *PARP1* and *PIK3CA* and genes unique for LUSC or LUAD (Table 1) and these genes could be proposed as potential separate prognostic factor for LUSC and LUAD. All of these genes are favorable for LUAD when their expression is lower than cutpoints, meaning that patients with increased expression of these genes will have poorer overall survival. However, in the vast majority of patients of studied cohort, most of these genes indicated lower expression level than cutpoints which might be related with lower effect of studied signaling pathway.

The inverse correlation of *KIF11* and *KIF4A* with overall survival in LUAD vs LUSC could be related with various regulation of mitosis and cellular transport between both of this subtypes. Abnormal kinesin expression could alter the equal distribution of genetic materials during cell mitosis leading to numerous defects in the daughter cells⁶⁹ thus targeting specific kinesins may create a strategy for differentiation of LUSC and LUAD. Similarly, different trends of overall survival in case of *CDC25A* and *CDK2* could indicate other course of cell cycle associated with disease progression that determines OS. Interestingly, similar level of *PIK3CA* cutpoint expression showed completely opposed effect for patients survival. *PIK3CA* is one of the PIK isoform taking part in signal transduction and is frequently mutated in cancer⁷⁰ however our analysis showed that its expression status might play role in predicting the different therapeutic effect in LUAD and LUSC. Moreover, gene such as *KIF2A*, due to the large number of patients in the cohort with increased (unfavorable) expression of this gene, should be considered as an important potential target in LUAD.

Interestingly, the results for DFS do not correspond in large extent with the results for OS. Our findings suggest that high expression of *BRCA1*, *BRCA2*, *KIF14*, *MCM5*, *MCM8*, *MSH2*, *PARP1* were unfavorable indicator of DFS in LUSC, whereas low expression of *E2F2*, *MAPK6*, *MAPK8* and high expression of *MCM2* were unfavorable indicator of DFS in LUAD (Table 2). This results suggest completely different mechanism of recurrence in these two lung cancer subtypes.

To summarize, the great potential of biomarker we proposed for being a prognostic factor is based on their biological background, primarily allegiance to biological processes frequently altered in cancer. However, qualifying patients to the favorable or unfavorable prognosis group based only on the determination of the cutpoint for a given gene is insufficient, as optimal cutpoints are usually dataset dependent thus could be unlikely to be the same in other studies. Moreover as single-variable relations do not take into account the values of all the other predictors, therefore further analyzes as well as in vitro research that could confirm our statements would be needed.

Compiling all data, it becomes evident that extremely significant differences exist between gene expression profiles of LUSC and LUAD concerning distinct downstream outcomes of various signalling pathways such as Notch, Wnt, Hh and ErbB. Overexpression of well-known oncogenes like *PIK3CA*, *MYC*, *RAD51* as well as genes involved in cell cycle and DNA repair, suggest that LUSC might have enhanced aggressiveness and migratory potential positively affects its predisposition to metastasis. Worth noting, the fact that LUSC is more common among “smokers” is not irrelevant for deregulation of these genes. The exposure to the DNA damaging factors contained in tobacco could be one of the causes of differences found between LUSC and LUAD. On the other hand, lower expression of DNA repair key players like *BRCA1*, *BRCA2*, *MSH2* and *MSH6* in LUAD could suggest that strategies of DNA repair in LUSC are very distinct from those in LUAD. The analyzed ontologies clearly differentiate LUSC from LUAD in terms of the expression, especially expression of particular genes. More importantly, further analysis have identified sets of genes that differently affect patients overall survival and disease free survival depending on tumor subtype and thus we conclude that they should be taken into account as separate set of potential prognostic markers for LUAD and LUSC. Our research provides direction for clinical treatment and molecular mechanism insight of differentiation between LUAD and LUSC that may help developing and assessing novel diagnostic and prognostic procedures for lung cancer.

The limitation of the present study might be the fact that there is no perfect way to validate the results. The publicly available resources offering RNAseq data are very limited; due to that reason to cross-validate and corroborate our findings, we have retrieved gene expression profiles from microarray platforms via GEO repository. Nevertheless, we succeeded to confirm the primary results to a large extent. On average, about 50% of genes that overlapped between TCGA and GEO-derived WGCNA modules showed the same trend in expression as well as partitioned patients according to LUSC/LUAD phenotype in PCA. It should be also stated that the observable remaining discrepancies between primary and cross-validation studies may have arisen for several reasons. Firstly, the normalization methods that apply to RNAseq and microarrays are very distinct regarding both, technology and data processing. Secondly, according to Tian F et al. RNAseq compared to microarray technology shows in general only 67–68% average reproducibility⁷¹. Finally, the cohorts itself may be biased or potential batch effect exists.

Summarizing, subtyping of LUAD and LUSC based on gene expression provides valuable information regarding differential biological mechanisms of cancer development and invasiveness reflected in clinical features of NSCLC tumours that stem from abrogated evolutionary signalling and its downstream outcomes.

Materials and methods

We obtained the RNA-Seq data of 515 LUAD and 501 LUSC cancer samples (RNAseq, level 3 RNASeqV2, RSEM normalized) and clinical data of the 522 LUAD and 501 LUSC patients from TCGA, downloaded from <http://gdac.broadinstitute.org/> (data status of Jan 28th, 2016). The present study included the analysis of the data that have been collected and processed by The Cancer Genome Atlas Research Network, therefore no approval of the institutional committee was required. The methods of biospecimen procurement, RNA isolation and RNA sequencing were previously described by The Cancer Genome Atlas Research Network^{28,72}. All experimental protocols were approved by a named institutional or licensing committee and the informed consent from all subjects or, if subjects were under 18, from a parent or legal guardian was obtained, as described therein. All methods were carried out in accordance with relevant guidelines and regulations.

The TCGA RNAseq data were combined with the patients' clinical outcome. Patients with missing any clinical or expression value were excluded from further analysis. Finally, we qualified a total of 515 LUAD and 499 LUSC samples. The summary of clinical characteristics of the both cohorts is shown in Supplementary Table 5.

To explore significant differences between patients with LUAD and LUSC we performed GSEA⁷³. Enrichment analysis was applied to 20,502 genes in terms of the Canonical Pathway database. Enrichment was subjected to GSEA by applying t-test with a weighted scoring scheme, default 1000 permutation and permutation type regarding the phenotype, using the default significance threshold of FDR < 0.25. To achieve the reproducibility of results we used the precise number in random seed parameter, which was 779,948,241.

According to GSEA results, we decided to focus on downstream target genes of four pathways: Notch, Hedgehog, Wnt and ErbB. Through the Gene Transcription Regulation Database (GTRD), available online at <http://gtrd.biouml.org/>^{74,75}, we made a list of targets of pathway-specific transcription factors, separately for each of the aforementioned pathways. We listed a total of 2949 downstream target genes of *HES1*, *HES2*, *HES4*, *HES7*, *HEY1*, *HEY2* and *HEYL*, 2981 targets genes of *GLI1*, *GLI2* and *GLI3*, 2571 target genes of *LEF1*, *TCF3* and *TCF4* and 5912 target genes of *Elk-1*, *c-Myc*, *c-Jun*, *c-Fos*, *STAT5A*, *STAT5B*, *FOXO1* for Notch, Hedgehog, Wnt and ErbB pathways, respectively.

Weighted gene co-expression networks were built using the WGCNA package in the R environment⁷⁶. A more detailed description can be found at <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/>. Briefly, pairwise Pearson's correlation matrix of expression was calculated and then transformed into an adjacency matrix. With function `pickSoftThreshold` we utilized soft-thresholding approach ($\beta = 6$ for Notch, Hh, Wnt pathway and $\beta = 8$ for ErbB pathway), to ensure a scale-free topology of the network with scale-free topology index (R_2) > 0.80. Then, we used the adjacency matrix to construct the topological overlap matrix (TOM), which corresponds to the overlap between pairs of interconnected genes. TOM was used to produce hierarchical clustering tree of genes, by `hclust` R function and "average" as a method. Genes sharing common expression profiles were clustered into modules by `dynamicTreeCut` algorithm with `minModuleSize = 30` and other parameters set to default. To identify modules that were significantly correlated with the trait of interest – a subtype of lung cancer, we calculated the correlation between ME and clinical trait. ME was considered as the first principal component of each gene module. For each module, we also defined module membership as a correlation between ME and gene expression.

Afterwards, GS was defined as the log₁₀ transformation of the p-value in the linear regression between gene expression and external trait. Besides, module significance (MS) was defined as the average GS for all the genes in a module.

The annotation of the gene ontology terms within WGCNA modules was performed through WGCNA-dedicated R packages such as `anRichment` and `anRichmentMethods` involving MSigDB repositories, i.e. C2 KEGG canonical pathways and C5 GO Biological Processes.

Furthermore, we performed spatial grouping of lung cancer patients through MFA according to various variables to determine the relevance of subtype (LUAD vs LUSC) and further associations with combined expression of genes derived from the most prominent WGCNA modules. Each module was treated as a separate group and the analysis was applied for each pathway separately. The MFA was applied using packages: `FactoMineR` and `factoextra`⁷⁷ within the R environment.

We performed a survival analysis using `Evaluate Cutpoints` application for R environment⁷⁸. Overall Survival (OS) and Disease Free Survival (DFS) was estimated by determination of optimal cutpoint splitting patients into two subgroups of favorable and unfavorable prognosis regarding expression of particular target genes as biomarker. In our analysis we used `cutp` algorithms of cutpoint determination in correlation with survival time and clinical outcome according to the following clinical parameters: "patient.person_neoplasm_cancer_status" and "patient.vital_status" as event indicator and "patient.days_to_last_followup" and "patient.days_to_death" as time of observation for DFS and OS, respectively.

To cross-validate data reliability and reproducibility, we extracted the microarray data of lung cancer patients from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>). The chosen series, GSE4573 and GSE12667, were based on the Affymetrix Human Genome U133A Array and Affymetrix Human Genome U133A Plus 2.0 Array platforms, respectively. The GSE4573 dataset submitted by Raponi et al. included 130 squamous cell lung carcinoma samples and 22,284 probesets⁷⁷, whereas GSE12667 submitted by Ding et al. included 75 lung adenocarcinomas samples and 54,676 probesets⁷⁸. For validation purposes, we extracted 22,277 probesets that overlapped between the Raponi's and Ding's projects and conducted WGCNA analogously to the primary analysis with soft thresholding power of $\beta = 20$. Several modules of initially identified through `dynamicTreeCut` algorithm tended to share very similar expression profiles, therefore we decided to merge them at height cut of 0.05, which corresponded to eigengenes correlation of 0.95. The annotation of ontological terms regarding C2 KEGG canonical pathways and C5 GO Biological Processes was additionally performed for all relevant modules of WGCNA. Subsequently, these findings were cross-validated with the primary results in three ways: 1) probes of the chosen,

significant modules were combined with the lists of pathway targets primarily considered to identify genes that overlapped between the results of TCGA and GEO data. Subsequently, logFC regarding the differential expression of the identified genes between LUSC and LUAD was calculated on TCGA and GEO data followed by a comparison of the expression trends referring to up- or downregulation (up: $\log_{2}FC > 0$, down: $\log_{2}FC < 0$) via Venn diagrams; 2) the analysis of ontological terms for each of WGCNA module was performed and compared with the list of initial annotation to identify commonly altered processes; and 3) dimensional grouping of LUSC and LUAD patients according to the resultant expression of genes that overlapped between each identified module and the list of pathway targets considered in the primary analysis by applying PCA with cancer subtype as qualitative supplementary variable.

Additionally, to confirm general trends regarding the gene expression in TCGA and GEO data, we extracted a total of 1000 genes with the highest logFC from our previously prepared TCGA data (20,502 genes) and cross-validated with Raponi's and Ding's datasets (13,513 genes). Moreover, we chose set of genes from the most significant WGCNA modules based on their differential expression profiles in LUSC and LUAD ($\log_{2}FC > 1.5$) from the primary analysis and compared with patterns of expression with GEO data. Validation results together with tables (Supplementary Tables 2–4) and figures (Supplementary Figs. 5–9) are described in Supplementary Material.

Full lists of GSEA results and derived datasets as well as the source codes for R analyses are publicly available at GitHub repository (<https://github.com/orzechmag/lungs>).

Data availability

Publicly available datasets were analysed in this study. These data can be found here: TCGA <http://gdac.broadinstitute.org/>; GEO <https://www.ncbi.nlm.nih.gov/geo/>.

Received: 5 August 2020; Accepted: 9 November 2020

Published online: 03 December 2020

References

- Torre, L. A., Siegel, R. L. & Jemal, A. Lung cancer statistics. *Adv. Exp. Med. Biol.* **893**, 1–19. https://doi.org/10.1007/978-3-319-24223-1_1 (2016).
- Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424. <https://doi.org/10.3322/caac.21492> (2018).
- Gironés, R., López, P., Chulvi, R., Cañabate, M. & Dolores, T. others Ten years of lung cancer in a single center: gender, histology, stage and survival. *J. Cancer Metastasis Treat.* **1**, 201–207 (2015).
- Galluzzo, P. & Bocchetta, M. Notch signaling in lung cancer. *Expert Rev. Anticancer Ther.* **11**, 533–540. <https://doi.org/10.1586/era.10.158> (2011).
- Tian, S. Classification and survival prediction for early-stage lung adenocarcinoma and squamous cell carcinoma patients. *Oncol Lett* **14**, 5464–5470. <https://doi.org/10.3892/ol.2017.6835> (2017).
- Travis, W. D. Pathology of lung cancer. *Clin. Chest Med.* **32**, 669–692. <https://doi.org/10.1016/j.ccm.2011.08.005> (2011).
- Cooper, W. A. *et al.* Expression and prognostic significance of cyclin B1 and cyclin A in non-small cell lung cancer. *Histopathology* **55**, 28–36 (2009).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674. <https://doi.org/10.1016/j.cell.2011.02.013> (2011).
- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70. [https://doi.org/10.1016/s0092-8674\(00\)81683-9](https://doi.org/10.1016/s0092-8674(00)81683-9) (2000).
- Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558. <https://doi.org/10.1126/science.1235122> (2013).
- Andersson, E. R., Sandberg, R. & Lendahl, U. Notch signaling: simplicity in design, versatility in function. *Development* **138**, 3593–3612. <https://doi.org/10.1242/dev.063610> (2011).
- Xu, K., Moghal, N. & Egan, S. E. Notch signaling in lung development and disease. *Adv. Exp. Med. Biol.* **727**, 89–98. https://doi.org/10.1007/978-1-4614-0899-4_7 (2012).
- De Langhe, S. P. & Reynolds, S. D. Wnt signaling in lung organogenesis. *Organogenesis* **4**, 100–108. <https://doi.org/10.4161/org.4.2.5856> (2008).
- Konigshoff, M. & Eickelberg, O. WNT signaling in lung disease: a failure or a regeneration signal?. *Am. J. Respir. Cell Mol. Biol.* **42**, 21–31. <https://doi.org/10.1165/rcmb.2008-0485TR> (2010).
- Velcheti, V. & Govindan, R. Hedgehog signaling pathway and lung cancer. *J. Thorac. Oncol.* **2**, 7–10. <https://doi.org/10.1097/JTO.0b013e31802c0276> (2007).
- Chen, H.-Y. *et al.* EGFR-activating mutations, DNA copy number abundance of ErbB family, and prognosis in lung adenocarcinoma. *Oncotarget* **7**, 9017–9025. <https://doi.org/10.18632/oncotarget.7029> (2016).
- Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594. <https://doi.org/10.4065/83.5.584> (2008).
- Zappa, C. & Mousa, S. A. Non-small cell lung cancer: current treatment and future advances. *Transl. Lung Cancer Res.* **5**, 288–300. <https://doi.org/10.21037/tlcr.2016.06.07> (2016).
- Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550. <https://doi.org/10.1038/nature13385> (2014).
- 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. <https://doi.org/10.1158/1078-0432.CCR-10-0199> (2010).
- Faruki, H. *et al.* Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. *J. Thorac. Oncol.* **12**, 943–953. <https://doi.org/10.1016/j.jtho.2017.03.010> (2017).
- Lucchetta, M. *et al.* Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response. *BMC Cancer* **19**, 824. <https://doi.org/10.1186/s12885-019-5965-x> (2019).
- Casimiro, M. C., Crosariol, M., Loro, E., Li, Z. & Pestell, R. G. Cyclins and cell cycle control in cancer and disease. *Genes Cancer* **3**, 649–657. <https://doi.org/10.1177/1947601913479022> (2012).
- Malumbres, M. & Carnero, A. Cell cycle deregulation: a common motif in cancer. *Prog. Cell Cycle Res.* **5**, 5–18 (2003).
- Burgess, J. T. *et al.* DNA repair pathways and their therapeutic potential in lung cancer. *Lung Cancer Manag.* **3**, 159–173 (2014).
- Dietlein, F., Thelen, L. & Reinhardt, H. C. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. *Trends Genet.* **30**, 326–339. <https://doi.org/10.1016/j.tig.2014.06.003> (2014).
- Iyama, T. & Wilson, D. M. 3rd. DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst.)* **12**, 620–636. <https://doi.org/10.1016/j.dnarep.2013.04.015> (2013).

28. Lee, M.-N. *et al.* Epigenetic inactivation of the chromosomal stability control genes BRCA1, BRCA2, and XRCC5 in non-small cell lung cancer. *Clin. Cancer Res.* **13**, 832–838. <https://doi.org/10.1158/1078-0432.CCR-05-2694> (2007).
29. Galimberti, F. *et al.* Targeting the cyclin E-Cdk-2 complex represses lung cancer growth by triggering anaphase catastrophe. *Clin. Cancer Res.* **16**, 109–120 (2010).
30. Fan, C. *et al.* Overexpression of major CDKN3 transcripts is associated with poor survival in lung adenocarcinoma. *Br. J. Cancer* **113**, 1735–1743. <https://doi.org/10.1038/bjc.2015.378> (2015).
31. Zang, X. *et al.* Identifying CDKN3 gene expression as a prognostic biomarker in lung adenocarcinoma via meta-analysis. *Cancer Inform.* **14**, 183–191. <https://doi.org/10.4137/CIN.S17287> (2015).
32. Soria, J. C. *et al.* Overexpression of cyclin B1 in early-stage non-small cell lung cancer and its clinical implication. *Cancer Res.* **60**, 4000–4004 (2000).
33. Yoshida, T., Tanaka, S., Mogi, A., Shitara, Y. & Kuwano, H. The clinical significance of Cyclin B1 and Wee1 expression in non-small-cell lung cancer. *Ann. Oncol.* **15**, 252–256. <https://doi.org/10.1093/annonc/mdh073> (2004).
34. Cangi, M. G. *et al.* Role of the Cdc25A phosphatase in human breast cancer. *J. Clin. Investig.* **106**, 753–761. <https://doi.org/10.1172/JCI19174> (2000).
35. Nishioka, K. *et al.* Clinical significance of CDC25A and CDC25B expression in squamous cell carcinomas of the oesophagus. *Br. J. Cancer* **85**, 412–421. <https://doi.org/10.1054/bjoc.2001.1934> (2001).
36. Sladek, T. E2F transcription factor action, regulation and possible role in human cancer. *Cell Prolif.* **30**, 97–105 (1997).
37. Chen, L., Yu, J. H., Lu, Z. H. & Zhang, W. E2F2 induction in related to cell proliferation and poor prognosis in non-small cell lung carcinoma. *Int. J. Clin. Exp. Pathol.* **8**, 10545–10554 (2015).
38. Gorgoulis, V. G. *et al.* Transcription factor E2F-1 acts as a growth-promoting factor and is associated with adverse prognosis in non-small cell lung carcinomas. *J. Pathol.* **198**, 142–156. <https://doi.org/10.1002/path.1121> (2002).
39. Imai, M. A., Oda, Y., Oda, M., Nakanishi, I. & Kawahara, E. Overexpression of E2F1 associated with LOH at RB locus and hyperphosphorylation of RB in non-small cell lung carcinoma. *J. Cancer Res. Clin. Oncol.* **130**, 320–326. <https://doi.org/10.1007/s00432-003-0538-3> (2004).
40. Park, S.-A. *et al.* E2F8 as a novel therapeutic target for lung cancer. *J. Natl. Cancer Inst.* <https://doi.org/10.1093/jnci/djv151> (2015).
41. Liu, X., Gong, H. & Huang, K. Oncogenic role of kinesin proteins and targeting kinesin therapy. *Cancer Sci.* **104**, 651–656 (2013).
42. Taniwaki, M. *et al.* Activation of KIF4A as a prognostic biomarker and therapeutic target for lung cancer. *Clin. Cancer Res.* **13**, 6624–6631. <https://doi.org/10.1158/1078-0432.CCR-07-1328> (2007).
43. Corson, T. W., Huang, A., Tsao, M.-S. & Gallie, B. L. KIF14 is a candidate oncogene in the 1q minimal region of genomic gain in multiple cancers. *Oncogene* **24**, 4741–4753. <https://doi.org/10.1038/sj.onc.1208641> (2005).
44. Zhou, X. *et al.* Prognostic value of KIF2A and TP53 overexpression in non-small cell lung cancer. *Int. J. Clin. Exp. Pathol.* **9**, 7266–7275 (2016).
45. Xie, T. *et al.* High KIF2A expression promotes proliferation, migration and predicts poor prognosis in lung adenocarcinoma. *Biochem. Biophys. Res. Commun.* **497**, 65–72. <https://doi.org/10.1016/j.bbrc.2018.02.020> (2018).
46. Schneider, M. A. *et al.* AURKA, DLGAP5, TPX2, KIF11 and CKAP5: five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. *Int. J. Oncol.* **50**, 365–372. <https://doi.org/10.3892/ijo.2017.3834> (2017).
47. Ha, S.-A. *et al.* Cancer-associated expression of minichromosome maintenance 3 gene in several human cancers and its involvement in tumorigenesis. *Clin. Cancer Res.* **10**, 8386–8395. <https://doi.org/10.1158/1078-0432.CCR-04-1029> (2004).
48. Kikuchi, J. *et al.* Minichromosome maintenance (MCM) protein 4 as a marker for proliferation and its clinical and clinicopathological significance in non-small cell lung cancer. *Lung Cancer* **72**, 229–237. <https://doi.org/10.1016/j.lungcan.2010.08.020> (2011).
49. Liu, Y.-Z. *et al.* MCMs expression in lung cancer: implication of prognostic significance. *J. Cancer* **8**, 3641–3647. <https://doi.org/10.7150/jca.20777> (2017).
50. Vigouroux, C. *et al.* Methyl(R217)HuR and MCM6 are inversely correlated and are prognostic markers in non small cell lung carcinoma. *Lung Cancer* **89**, 189–196. <https://doi.org/10.1016/j.lungcan.2015.05.008> (2015).
51. Yang, J. *et al.* Prognostic significance of MCM2, Ki-67 and gelsolin in non-small cell lung cancer. *BMC Cancer* **6**, 203. <https://doi.org/10.1186/1471-2407-6-203> (2006).
52. Mitani, S. *et al.* Analysis of c-myc DNA amplification in non-small cell lung carcinoma in comparison with small cell lung carcinoma using polymerase chain reaction. *Clin. Exp. Med.* **1**, 105–111. <https://doi.org/10.1007/s10238-001-8020-5> (2001).
53. Van Waardenburg, R. C. *et al.* Effects of an inducible anti-sense c-myc gene transfer in a drug-resistant human small-cell-lung-carcinoma cell line. *Int. J. Cancer* **73**, 544–550. [https://doi.org/10.1002/\(sici\)1097-0215\(19971114\)73:4%3c544::aid-ijc15%3e3.0.co;2-9](https://doi.org/10.1002/(sici)1097-0215(19971114)73:4%3c544::aid-ijc15%3e3.0.co;2-9) (1997).
54. Hibi, K. *et al.* AIS is an oncogene amplified in squamous cell carcinoma. *Proc. Natl. Acad. Sci. USA* **97**, 5462–5467. <https://doi.org/10.1073/pnas.97.10.5462> (2000).
55. Pelosi, G. *et al.* p63 immunoreactivity in lung cancer: yet another player in the development of squamous cell carcinomas?. *J. Pathol.* **198**, 100–109. <https://doi.org/10.1002/path.1166> (2002).
56. Zeng, J. *et al.* GSK3beta overexpression indicates poor prognosis and its inhibition reduces cell proliferation and survival of non-small cell lung cancer cells. *PLoS ONE* **9**, e91231. <https://doi.org/10.1371/journal.pone.0091231> (2014).
57. Scheffler, M. *et al.* PIK3CA mutations in non-small cell lung cancer (NSCLC): genetic heterogeneity, prognostic impact and incidence of prior malignancies. *Oncotarget* **6**, 1315–1326. <https://doi.org/10.18632/oncotarget.2834> (2015).
58. Zhu, Y.-F. *et al.* PI3K expression and PIK3CA mutations are related to colorectal cancer metastases. *World J. Gastroenterol.* **18**, 3745–3751. <https://doi.org/10.3748/wjg.v18.i28.3745> (2012).
59. Akagi, I. *et al.* Overexpression of PIK3CA is associated with lymph node metastasis in esophageal squamous cell carcinoma. *Int. J. Oncol.* **34**, 767–775. https://doi.org/10.3892/ijo_00000202 (2009).
60. Aleskandarany, M. A. *et al.* PIK3CA expression in invasive breast cancer: a biomarker of poor prognosis. *Breast Cancer Res. Treat.* **122**, 45–53. <https://doi.org/10.1007/s10549-009-0508-9> (2010).
61. Long, W. *et al.* ERK3 signals through SRC-3 coactivator to promote human lung cancer cell invasion. *J. Clin. Investig.* **122**, 1869–1880. <https://doi.org/10.1172/JCI61492> (2012).
62. Takahashi, H., Ogata, H., Nishigaki, R., Broide, D. H. & Karin, M. Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta- and. *Cancer Cell* **17**, 89–97. <https://doi.org/10.1016/j.ccr.2009.12.008> (2010).
63. Zhao, X. *et al.* Cell cycle-dependent control of homologous recombination. *Acta Biochim. Biophys. Sin. (Shanghai)* **49**, 655–668. <https://doi.org/10.1093/abbs/gmx055> (2017).
64. Ko, J.-C. *et al.* Role of repair protein Rad51 in regulating the response to gefitinib in human non-small cell lung cancer cells. *Mol. Cancer Ther.* **7**, 3632–3641. <https://doi.org/10.1158/1535-7163.MCT-08-0578> (2008).
65. Qiao, G.-B. *et al.* High-level expression of Rad51 is an independent prognostic marker of survival in non-small-cell lung cancer patients. *Br. J. Cancer* **93**, 137–143. <https://doi.org/10.1038/sj.bjc.6602665> (2005).
66. Chen, K. *et al.* An analysis of the gene interaction networks identifying the role of PARP1 in metastasis of non-small cell lung cancer. *Oncotarget* **8**, 87263–87275. <https://doi.org/10.18632/oncotarget.20256> (2017).
67. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98. <https://doi.org/10.1038/cr.2007.115> (2008).
68. Downey, C. M. & Jirik, F. R. DNA mismatch repair deficiency accelerates lung neoplasm development in. *Cancer Med.* **4**, 897–902. <https://doi.org/10.1002/cam4.420> (2015).

69. Yue, Y. & Yu-Mei, F. The role of kinesin family proteins in tumorigenesis and progression: potential biomarkers and molecular targets for cancer therapy. *Cancer* **116**(22), 5150–5160. <https://doi.org/10.1002/cncr.25461> (2010).
70. Alqahtani, A., Ayesh, H. S. K. & Halawani, H. PIK3CA gene mutations in solid malignancies: association with clinicopathological parameters and prognosis. *Cancers (Basel)* **12**(1), 93. <https://doi.org/10.3390/cancers12010093> (2019).
71. Tian, F., Wang, Y., Seiler, M. & Hu, Z. Functional characterization of breast cancer using pathway profiles. *BMC Med Genom.* **7**, 45. <https://doi.org/10.1186/1755-8794-7-45> (2014).
72. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **2012**(489), 519–525. <https://doi.org/10.1038/nature11404> (2012).
73. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550. <https://doi.org/10.1073/pnas.0506580102> (2005).
74. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucl. Acids Res.* **45**, D61–D67. <https://doi.org/10.1093/nar/gkw951> (2017).
75. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation-2019 update. *Nucl. Acids Res.* **47**, D100–D105. <https://doi.org/10.1093/nar/gky1128> (2019).
76. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. <https://doi.org/10.1186/1471-2105-9-559> (2008).
77. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw. Art.* **25**, 1–18. <https://doi.org/10.18637/jss.v025.i01> (2008).
78. Oguszk, M., Orzechowska, M., Jędraszka, D., Witas, P. & Bednarek, A. K. Evaluate cutpoints: adaptable continuous data distribution system for determining survival in Kaplan–Meier estimator. *Comput. Methods Programs Biomed.* **177**, 133–139. <https://doi.org/10.1016/j.cmpb.2019.05.023> (2019).
79. Raponi, M. *et al.* Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.* **66**, 7466–7472. <https://doi.org/10.1158/0008-5472.CAN-06-1191> (2006).
80. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075. <https://doi.org/10.1038/nature07423> (2008).

Acknowledgments

The results shown here are in whole based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

Author contributions

A.K.B. was responsible for study conceptualization, funding acquisition, project administration, supervision, reviewing and editing the manuscript. D.A. & M.O. were responsible for data curation, investigations and analyses, visualization and interpretation of the findings, validation of the results and writing the original draft of the manuscript. All authors have approved the submitted version of the manuscript and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Funding

This research was funded by the National Science Centre, Poland, grant number 2016/23/N/NZ5/02575 and Medical University of Lodz, grant number 503/0-078-02/503-01-004.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-77284-8>.

Correspondence and requests for materials should be addressed to D.A.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020