Progressive lung cancer determined by expression profiling and transcriptional regulation

NAMSHIK HAN[1,2*], ZULKIFLI DOL[1*], OLGA VASIEVA[3], RUSSELL HYDE[3,4], TRIANTAFILLOS LILOGLOU[4], OLAIDE RAJI[4], ELISABETH BRAMBILLA[5], CHRISTIAN BRAMBILLA[5], YVES MARTINET[6], GABRIELLA SOZZI[7], ANGELA RISCH[8], LUIS M. MONTUENGA[9], THE EUELC CONSORTIUM, ANDY BRASS[1,2] and JOHN K. FIELD[4]

[1]School of Computer Science, The University of Manchester, Kilburn Building, Oxford Road; [2]Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Dover Street, Manchester; [3]Institute of Integrative Biology, University of Liverpool; [4]Roy Castle Lung Cancer Research Programme, University of Liverpool Cancer Research Centre, Department of Clinical and Molecular Cancer Medicine, Liverpool, UK; [5]Albert Bonniot Institute, INSERM U823, Joseph Fourier University Grenoble; [6]Centeral Hospitalier Universitaire de Nancy, France; [7]Department of Experimental Oncology, Milan, Italy; [8]German Cancer Research Centre, Heidelberg, Germany; [9]Center for Applied Medical Research, CIMA, University of Navarra, Pamplona, Spain

Received January 9, 2012; Accepted February 10, 2012

DOI: 10.3892/ijo.2012.1421

Abstract. Clinically, our ability to predict disease outcome for patients with early stage lung cancer is currently poor. To address this issue, tumour specimens were collected at surgery from non-small cell lung cancer (NSCLC) patients as part of the European Early Lung Cancer (EUELC) consortium. The patients were followed-up for three years post-surgery and patients who suffered progressive disease (PD, tumour recurrence, metastasis or a second primary) or remained disease-free (DF) during follow-up were identified. RNA from both tumour and adjacent-normal lung tissue was extracted from patients and subjected to microarray expression profiling. These samples included 36 adenocarcinomas and 23 squamous cell carcinomas from both PD and DF patients. The microarray data was subject to a series of systematic bioinformatics analyses at gene, network and transcription factor levels. The focus of these analyses was 2-fold: firstly to determine whether there were specific biomarkers capable of differentiating between PD and DF patients, and secondly, to identify molecular networks which may contribute to the progressive tumour phenotype. The experimental design and analyses performed permitted the clear differentiation between PD and DF patients using a set of biomarkers implicated in neuroendocrine signalling and allowed the inference of a set of transcription factors whose activity may differ according to disease outcome. Potential links between the biomarkers, the transcription factors and the genes p21/CDKN1A and Myc, which have previously been implicated in NSCLC development, were revealed by a combination of pathway analysis and microarray meta-analysis. These findings suggest that neuroendocrine-related genes, potentially driven through p21/CDKN1A and Myc, are closely linked to whether or not a NSCLC patient will have poor clinical outcome.

Correspondence to: Professor John K. Field, Roy Castle Lung Cancer Research Programme, University of Liverpool Cancer Research Centre, Department of Clinical and Molecular Cancer Medicine, University of Liverpool, 200 London Road, Liverpool, L3 9TA, UK
E-mail: j.k.field@liv.ac.uk

*Contributed equally

Key words: lung cancer, microarray, network analysis, cancer progression

Introduction

Cancer of the lung kills more patients than any other cancer worldwide. In 2008 in England and Wales, it accounted for 24% of all male cancer deaths and 20% of all female cancer deaths (1), representing 6% of all deaths in the UK. The two main clinically relevant subtypes of lung cancer are small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter can be histologically subclassified into adenocarcinoma, squamous cell carcinoma and large-cell carcinoma (2).

The European Early Lung Cancer (EUELC) consortium (3) comprises 12 centres in eight European countries: France, Germany, Ireland, Italy, the Netherlands, Poland, Spain and the UK. Through the EUELC, more than 900 non-small cell lung cancer (NSCLC) patients were recruited from around Europe, and their specimens were stored at the European Bronchial Tissue Bank based at the Roy Castle Lung Cancer Research Programme in Liverpool. The NSCLC patients were followed-up for 36 months.

One of the aims of the consortium was to ascertain whether alterations in gene expression caused by lung carcinogenesis are detectable at an early stage in the respiratory epithelium and whether these changes can be used to predict disease outcome. To this end, the consortium developed a nested case/control study to search for molecular-pathological differences seen in

those patients whose cancer was seen to re-occur (PD, progressive disease) and those who remained disease-free (DF) in the study period. The PD classification included patients in which tumour recurrence, metastasis, or a second primary tumour was observed (3). It was hypothesized that expression profiling of the RNA from these tumours may identify an expression signature associated with lung cancer patients with a poor disease outcome.

Tissue was taken from patients after surgical resection for lung cancer (3). The tumour specimens included both cancer cells and non-cancer cells from adjacent tissue. RNA from cancer cells and adjacent normal cells was isolated, enabling comparisons between expression profiles of cancer and adjacent normal lung tissue. This allowed one to control for field effects (such as inflammatory signals) that are typical of the lung tumour environment, but not intrinsic to transformed lung cells. The PD patients were matched to DF patients based on follow-up time (at least as long as the event time of PD subjects), centre, gender, age (± 6 years), and histology/nodal stage. In total this EUELC genome wide expression profiling dataset contains 59 patients with 41,672 candidate transcripts. It is from this dataset that we have looked for markers able to distinguish between patients who remained disease-free compared with those that developed a recurrence of the cancer.

A number of studies have used tumour tissue to predict post-surgery disease outcome, either based on histology (4,5), microarray data (6-9), or the expression levels of specific genes, proteins, and immune markers (10-12). In particular, molecules such as neurone-specific enolase (13) and p21/CDKN1A (14,15) have been associated with poor prognosis. Moreover, several studies have identified markers of neuroendocrine differentiation in NSCLC and have associated this molecular phenotype to increased tumour recurrence (reviewed in ref. 16). Indeed, the frequent presence of neuroendocrine markers in NSCLC samples has led to discussions regarding how such cancers should be classified (17).

In this study we have chosen to take an agnostic approach to the analysis, using a combination of machine learning, bioinformatics and pathway analysis, essentially a systems strategy, to examine the EUELC expression dataset. The aim was to identify markers of disease outcome in the data, using machine learning and text mining, that relate to biological processes, discovered through network and pathway analysis, which may be characteristic of a progressive tumour phenotype and potentially useful for the clinical management of patients.

Methods and Results

i) Trancriptional analysis and biomarker discovery

Transcriptional analysis. Tumour samples were taken from patients from the EUELC cohort which had undergone curative surgery. High quality RNA (Agilent - RIN > 6), from frozen tumour samples was profiled by Oxford Gene Technology (3). Normal samples were taken from pooled samples of tissue proximal to the tumours. Total RNA (1 μg) was labeled using low input RNA Amp kit (Agilent 5184-3523). Lung cancer RNAs were labeled with Cy3 and were hybridised with a Cy5 reference sample labeled from RNA consisting of a 50:50 mix of Universal human reference (Stratagene 740000) and a Human lung cell line (Ambion AM7864). The labeled samples were then hybridised to Agilent 44k human whole genome oligo microarrays (Agilent G4112A).

EUELC collaboration pathologists classified the tumours as being of either squamous cell carcinoma or adenocarcinoma origin. Patient records were then used to determine whether these were from patients who remained disease-free (DF) or suffered a cancer recurrence (PD, progressive disease) over the 3-year follow-up period. There were therefore 8 classes of data within the study, represented as: adenocarcinoma progressive disease cancer (A^c_{pd}); adenocarcinoma disease-free cancer (A^c_{df}); squamous cell carcinoma progressive disease cancer (S^c_{pd}); squamous cell carcinoma disease-free cancer (S^c_{df}); and the adjacent normals to each sample (A^n_{pd}, A^n_{df}, S^n_{pd}, S^n_{df}). The numbers of distinct samples in each group is shown in Table I.

The microarray data was normalized using a standard Lowess normalisation (18) in the MaxD software (19) to produce a set of log expression data. Within the data there were a significant number of missing values across the complete data set (~5%) affecting more than 2000 probes. The missing values were imputed using a standard missing value imputation [mean imputation (20)].

Log ratios of expression values between the cancers and adjacent normal groups were evaluated by subtracting the log of the expression values of the adjacent normal tissue from each of the appropriate log cancer expression datasets. This generated 4 new datasets: $\log(A_{pd}) = \log(A^c_{pd}) - \log(A^n_{pd})$; $\log(A_{df}) = \log(A^c_{df}) - \log(A^n_{df})$; $\log(S_{pd}) = \log(S^c_{pd}) - \log(S^n_{pd})$; $\log(S_{df}) = \log(S^c_{df}) - \log(S^n_{df})$. Each of these datasets represents the differences in gene expression between each type of cancer in the dataset and the adjacent normal tissue to that cancer.

In order to gain a deeper understanding of the data we performed a principal components analysis (PCA) of the data, projecting the experimental data into a low dimensional vector space. The PCA plot of the data is shown in Fig. 1. It is particularly striking that the data can be clearly separated in two distinct ways: i) based on tumour histology (squamous or adenocarcinoma); and ii) based on disease outcome - that is, whether the cancer that will recur (PD) or not (DF) within 3 years of surgery. This suggests that it should be possible to identify markers within the dataset which will allow us to predict whether a given cancer belongs to the PD or DF groups.

Biomarker discovery. Biomarkers selection technique was performed using a standard machine learning strategy. Five separate algorithms were used to identify genes that differentiated between the PD and DF classes using a leave-one-out cross-validation strategy. Each method provided a ranking of the genes as to their ability to separate samples between the two classes based on different assumptions as to the underlying nature of the data. A majority rule strategy was then used to select the set of genes which performed consistently across all 5 methods. This analysis provided a set of the 23 genes selected (shown in Fig. 2). A heat map analysis (Fig. 2) clearly shows that these genes do provide a mechanism for separating the two classes of patient data (DF or PD) and that these genes appear to function similarly in patients classified as having a squamous or adenocarcinoma phenotype.

A simple pathway analysis was performed (enrichment for particular gene pathways in the KEGG database performed using the DAVID software). This analysis showed a very significant enrichment for genes linked to neuroendocrine path-

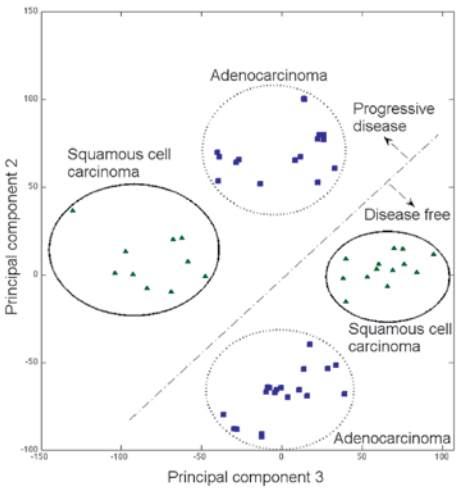

Figure 1. Principal components of the normalised microarray dataset. Second and third principal components separate lung tumours from patients exhibited progressive disease, or remained disease-free, during follow-up. SCC, squamous cell carcinoma.

ways (specifically RIT2, MTNR1A, GNG3, CNTN2, VHLL, PTHR2, GPR50 and TPH2) - which was surprising given the small number of genes identified in the biomarker set from the machine learning analysis.

ii) Network and pathway analysis

Network analysis. The biomarker analysis above provides some insights into the cancer processes. However, if we are to understand what might be driving these changes it would be desirable to understand the fundamentals of regulatory machinery, whether there are any hints as to changes in the transcriptional regulatory machinery between DF and PD patient samples. We have therefore performed comprehensive analyses of the transcriptional machinery using some recently developed theoretical tools (21). We consider that there are a number of components involved in transcription regulation: a) a set of genes and associated transcription factor binding sites (*cis*-regulatory regions); b) a set of TF proteins that bind to the *cis*-regulatory regions to regulate gene expression; c) a network

Figure 2. Gene expression profiles for genes which distinguish post-operative prognosis. Heat map of the four lung cancer classes considered: adenocarcinoma with progressive disease; adenocarcinoma with disease-free follow-up period; squamous cell carcinoma with progressive disease; squamous cell carcinoma with disease-free follow-up period.

Table I. EUELC PD & DF specimens selected for expression profiling

Adenocarcinoma				Squamous cell carcinoma			
Progressive disease		Disease-free		Progressive disease		Disease-free	
Normal	Cancer	Normal	Cancer	Normal	Cancer	Normal	Cancer
3	17	3	19	2	10	2	13

B, Tumour specimens PD v DF matching, pathology, RIN and pTNM staging

Adenocarcinoma				Squamous cell carcinoma			
Patient ID	PD matching	RNA/RIN	pTNM	Patient ID	PD matching	RNA/RIN	pTNM
A-002		8.5	T2N0	B-032		5	T2N0
A-041	A-002	7	T1N0	B-015	B-032	6.9	T2N0
B-025		9	T1N0	B-043		7.1	T2N1
B-031	B-025	6.1	T1N0	B-021	B-043	5.5	T1N1
B-042	B-025	8.8	T2N0	B-005	B-043	6.3	T2N1
B-035		6.4	T2N0	C-053		7.1	T2N0
B-018	B-035	7	T2N0	C-009	C-053	6.8	T2N0
B-022	B-035	6.7	T1N0				
				E-031		7.4	T2N0
B-048		6.1	T2N0	E-034	E-031	8.2	T2N0
B-047	B-048	6.4	T1N0	E-069	E-031	6.2	T2N0
B-009	B-048	6.5	T1N0				
				F-008		5.1	T2N0
C-019		5	T2N0	F-046	F-008	7.4	T1N0
C-068	C-019	7.5	T1N0				
				B-003		7.6	T2N1
D-028		6.4	T2N0	B-027	B-003	9.2	T1N1
D-123	D-028	6.8	T1N0				
				B-029		9.3	T2N1
D-033		7.2	T2N0	B-006	B-032	8.8	T2N0
D-001	D-033	6	T1N0				
				B-050		N/A	T3N1
D-067		6.9	T1N0	B-033	B-043	8.7	T2N1
D-059	D-067	7	T1N0				
D-069	D-067	6.6	T1N0	C-022		7.1	T2N0
				C-030	C-002	7.1	T1N0
E-042		5.1	T2N0				
E-041	E-042	6.4	T2N0	D-131		8.5	T3N1
E-046	E-042	6.8	T2N0	D-096	D-131	7.8	T2N1
D-105		7	T2N1	F-062		9.5	T2N1
D-077	D-105	6.7	T1N1	F-030	F-062	5	T1N1
01-016		7.8	T2N0				
01-011	01-016	8.4	T1N0				

Table I. Continued.
C, Normal pooled samples for PD and DF

Normal pool ID	Patient ID	PD matching	PD/DF	Gender	Patient diagnosis	RIN
PD1	C-028	C-028	PD	Male	Adenocarcinoma	6.8
PD1	C-075	C-075	PD	Male	Adenocarcinoma	7.6
PD1	A-025	A-025	PD	Male	Adenocarcinoma	8.9
PD2	B-048	B-048	PD	Female	Adenocarcinoma	6.7
PD2	C-059	C-059	PD	Male	Adenocarcinoma	7.4
PD2	C-049	C-049	PD	Male	Adenocarcinoma	7.2
PD3	F-062	F-062	PD	Male	SCC	6
PD3	G-024	G-024	PD	Male	SCC	7
PD3	C-069	C-069	PD	Male	SCC	5.7
PD4	B-043	B-043	PD	Male	SCC	6.8
PD4	C-053	C-053	PD	Male	SCC	7.8
PD4	C-033	C-033	PD	Male	SCC	7.2
DF1	B-009	B-079	DF	Female	Adenocarcinoma	7
DF1	B-047	B-048	DF	Female	Adenocarcinoma	7.3
DF1	D-001	D-033	DF	Male	Adenocarcinoma	6
DF2	C-087	C-113	DF	Male	Adenocarcinoma	6.2
DF2	D-070	D-033	DF	Male	Adenocarcinoma	5.3
DF2	C-060	C-109	DF	Male	Adenocarcinoma	7.6
DF3	E-034	E-031	DF	Male	SCC	7
DF3	D-065	D-057	DF	Male	SCC	7.2
DF3	B-005	B-003	DF	Male	SCC	6.8
DF4	E-069	E-031	DF	Male	SCC	8.1
DF4	C-022	C-002	DF	Male	SCC	5.5
DF4	C-030	C-002	DF	Male	SCC	6.1

describing the interactions between genes and transcription factors; and d) the measured set of gene expression profiles. The inputs of the circuits (a and b) are processed on the circuit architecture (c) to generate the output (d).

A mathematical model (22) which attempts to describe the relations of the four components in the transcriptional regulatory circuits was applied to infer the TF activity and concentration levels. We can represent the log gene expression measurements (the output of the regulatory circuits) as a vector $\underline{\varepsilon}$ where each element of the vector represents the signal from a different gene. The connection between TFs and genes can be represented as a binary matrix $\underline{\underline{\tau}}$ in which the rows and columns represent the genes and TFs (connection topology). $\underline{\underline{\alpha}}$ is a matrix that captures the strength of the interaction between TFs for their target gene. $\underline{\lambda}$ is the vector of concentrations of each of the TFs (the input: TFs), and $\underline{\nu}$ is a vector of independent and identically distributed variables modeling the noise in the system. We can therefore model gene expression in the form: $\underline{\varepsilon} = \underline{\underline{\alpha}}\underline{\underline{\tau}}\underline{\lambda} + \underline{\nu}$.

Given a knowledge of *cis*-regulatory regions $\underline{\underline{\tau}}$ and gene expression results $\underline{\varepsilon}$, the model can be used to infer $\underline{\underline{\alpha}}$ and $\underline{\lambda}$ - giving us a predictive model of the underlying TF concentrations and interactions underlying the measured response. Differences in these between DF and PD patients can provide useful insights into the differences in TF activity in patients whose cancer does or does not recur. In particular we are interested in situations in which the activity of a TF changes, not just because it changes in concentration, but as a result of a change in its state - for example because of a change in the cofactors or phosphorylation state. To capture this we look at the ratio between the activity level of the TFs ($\underline{\alpha}$) and the concentration, $\underline{\lambda}$. We therefore define a new variable, the scaled total interaction, for each transcription factor, *f*, that sums its total effect across all the genes with which it interacts, scaled by the its concentration:

$$\sigma_f = \sum_{g=1}^{n} \left| \frac{\alpha_{g,f}}{\lambda_f} \right|$$

The total scaled interaction can be calculated for every TF in the dataset in each of the cancer datasets. In particular we can look at this factor as a function of whether a tissue is from the normal or cancer set, this provides information on TFs have different activation states between the cancer and normal tissues. Similarly we can look the set of data from the DF and PD populations. In particular we calculated two new parameters for each TF: $f_x = \sigma_f(cancer) - \sigma_f(normal)$ and $f_y = \sigma_f(PD) - \sigma_f(DF)$.

By plotting these two variables we can easily identify TFs that change activation in cancer versus normal tissues and

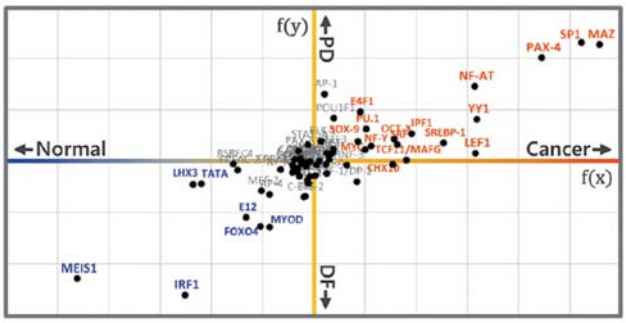

Figure 3. Global view of transcription factor activities in both adenocarcinoma and squamous cell carcinoma. TF-activity changes between cancer and normal tissue (x-axis) and between PD and DF tumours (y-axis) are indicated. Origin-proximal TFs are unaffected by either cancer or patient group. TFs in the upper-right upregulate target genes in cancer cells and PD patients, those in the lower left upregulate their targets in normal cells and DF patients.

between the PD and DF patients. The results of this analysis are presented in Fig. 3.

The computational pipeline, therefore, has enabled the analysis of TFs which appear to be strongly related to the progressive disease versus disease-free phenotypes. Fig. 3 provides a global view of the TFs responsive levels for the lung cancer as it presents all changes across two types of cells (TF states) and two groups of patients (TF preferences) in a single shot image. The major TFs which have been up-regulated in the cancer cells are MAZ, SP1, PAX4, NF-AT, YY1, LEF1. The strongly activated TFs in the normal cells included MEIS1, IRF1 (Fig. 3). Interestingly, the up-regulated TFs in the cancer cells have preference to the progressive disease phenotypes. In contrast, the active TFs in normal cells are found in disease-free phenotypes.

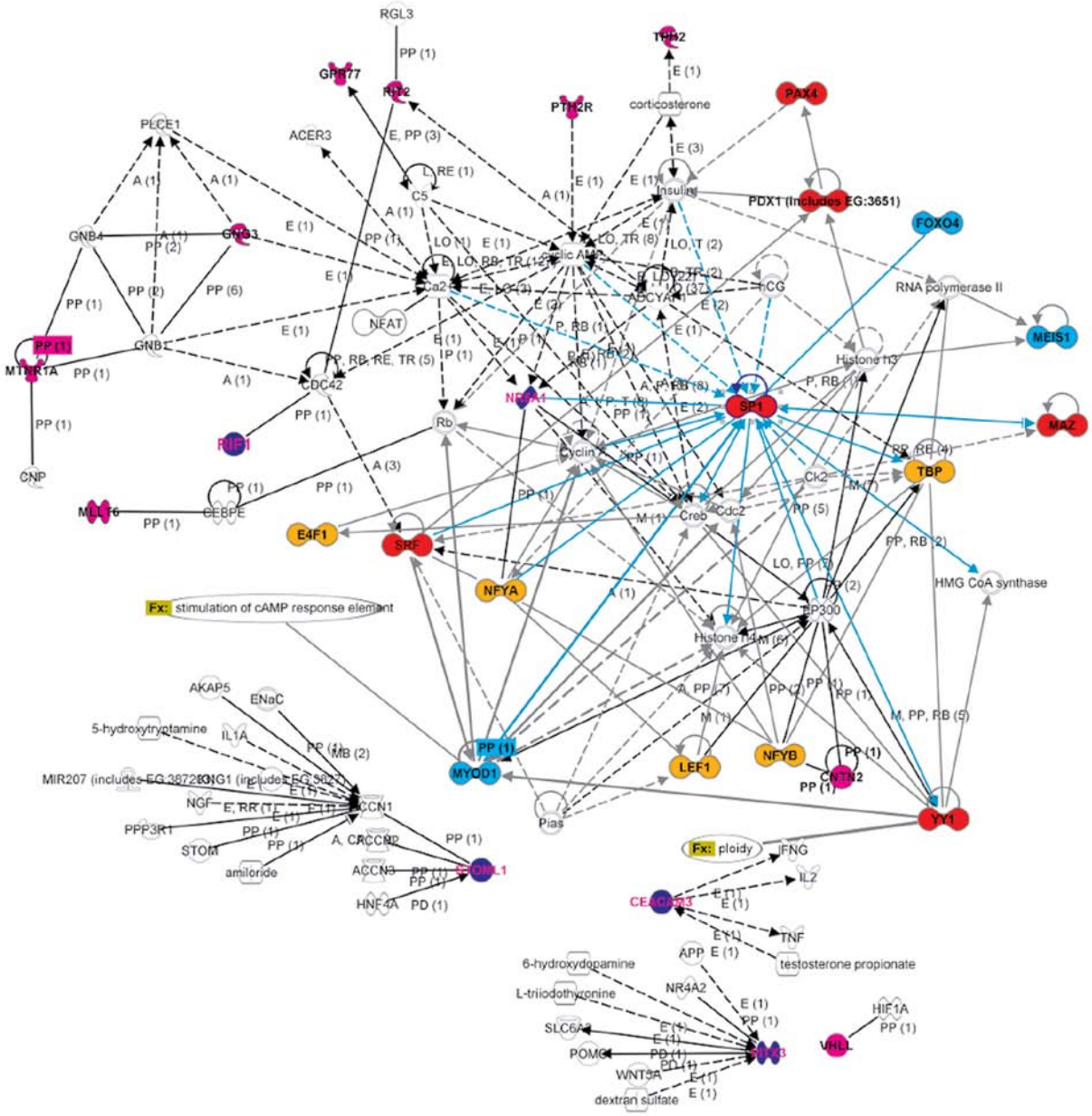

Figure 4. Functional interaction network for activated transcription factors and differentially expressed transcripts in PD. Network nodes are coloured according to gene/protein groupings: functional markers (pink, PD-upregulated; dark blue, PD-downregulated); predicted TFs (red/orange, high/moderate activation in PD; light blue, suppressed in PD). Connections of first- and second-order with cAMP are shown in dark and light blue, respectively. White nodes were absent from the input lists but included by the network constructing software.

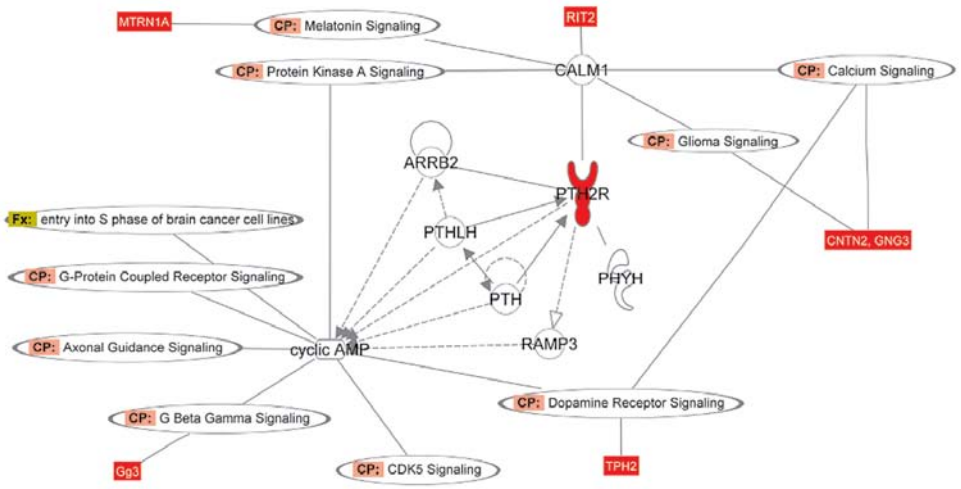

Figure 5. Connections between differentially-expressed PD markers and ontologies characteristic of neuronal function. Details of a functional connection between PTH2R and cAMP, and cross-talk between PD differential markers (red figures) are shown.

Methods for pathway analysis. To study a relationship between the revealed functions we used Ingenuity Pathway Analysis (IPA) (23,24) software. The biological information in IPA is created by text mining tools, and the data retrieved from the major protein-protein interaction databases and the literature can be used to construct networks based on functional connectivity between different molecules. IPA also allows the direct projection of physiologically and pathologically relevant data onto biological networks and pathways.

Links between biomarkers and known biological information. To study the relationship between the predicted TFs and known biological functions, we constructed a list of HUGO gene symbols corresponding to up- and down-regulated differential PD markers and predicted highly up-regulated TFs. This list was used as input to the Ingenuity database of protein and metabolic interactions from which a network comprising the marker functions and the predicted TFs was generated (Fig. 4). IPA software included a few additional nodes to construct a continuous network by connecting otherwise isolated parts of the dataset. The process of connectivity reconstruction is entirely automatic and thus helps to reveal functionally important hubs that may be central to an integral network. Ca^{2+}, cAMP-response module, histones 3 and 4, P300 and a group of cell cycle regulating functions present such additional nodes in a network that connects PD marker functions to the predicted TFs, with only 7 from 50 integrated functions not being connected to a central cAMP hub by either the first or the second order interactions. Fig. 5 shows details of a functional connection between PTHR2 and cAMP. It also schematically shows cross-talk between markers for progression (red boxes) and demonstrates their relation to neuronal physiology.

When marker functions were combined with all predicted TFs as input to Ingenuity, a c-myc centred network was obtained (Fig. 6). Meta-analysis of previously published expression data using Genevestigator software (25) identified a correlation between the expression of PD differential markers and the presence of active Myc. Fig. 7 demonstrates average relative levels of expression of the chosen PD markers across a variety of conditions. As can be seen, c-myc depletion causes the activation of almost all of the PD functional gene markers. Conversely, p73 overexpression leads to the downregulation of PD markers.

Transcription factor network. TFs enriched in the promoters for those genes which were differentially expressed in PD were identified. To discern whether functional relationships between these proteins exist, the corresponding gene symbols were used as input to Ingenuity. All of the TFs had interactions of first order with at least one other member of the group or with the TF Myc. Notably, from the 18 TF considered 10 had a direct functional relationship with Myc (Fig. 6).

A subset of 10 TFs that were highly enriched for the promoters of the same PD-regulated genes leads to another smaller network centred around CDKN1A/p21 (Fig. 8). HUGO symbols for the 10 selected TFs and all of the protein-protein, functional and genetic interactions in the Ingenuity database were used to automatically construct this network. Within the set of 10 TFs, 7 had a direct functional link to CDKN1A/p21.

Discussion

In this study, tumour expression profiles were used to characterise the 3-year cancer-free survival of lung cancer patients after surgery. System-wide analysis of mRNA expression in the context of transcriptional and regulatory networks identified several candidate markers for tumour progression and potential mechanisms for their persistent upregulation. This hierarchical application of bioinformatics to modelling tumour expression profiles highlighted a role for neuroendocrine signalling in progressive adenocarcinoma and squamous cell carcinoma of the lung.

Neuroendocrine genes as markers of progressive tumours. The nervous system controls physiological events through the neuroendocrine system, a network of hormone releasing glands. In the lung, pulmonary neuroendocrine cells are involved in lung morphogenesis and the serotonin-mediated response to hypoxia (26). Neuroendocrine tumours derived from these cells,

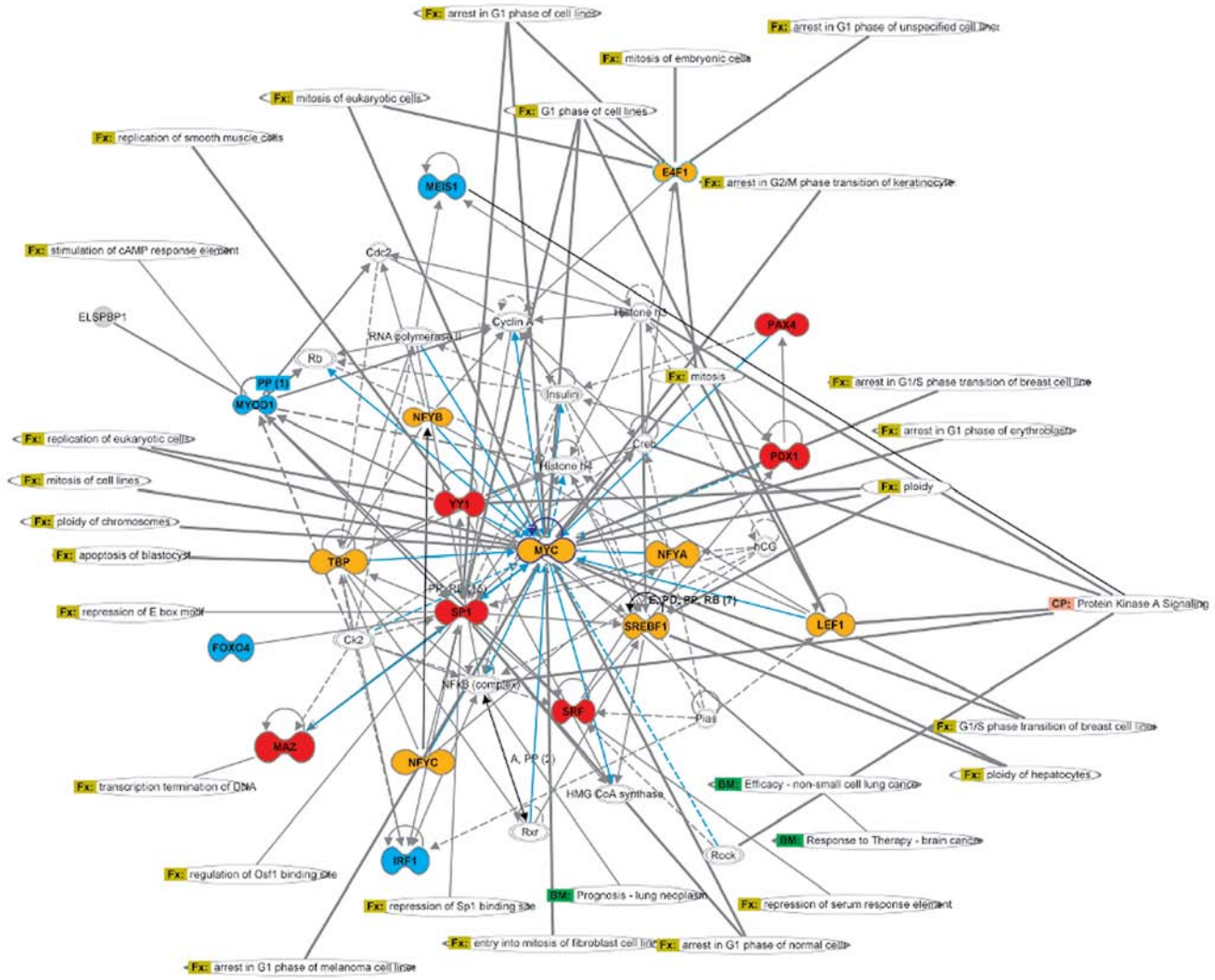

Figure 6. Myc-centred transcription factor network. Network linking the transcription factors with the greatest differential activity in PD tumours (red/orange, highly/moderately activated TFs; blue, suppressed TFs). First-order connections to Myc are shown in blue. Functional categories and selected ontologies are projected onto the network.

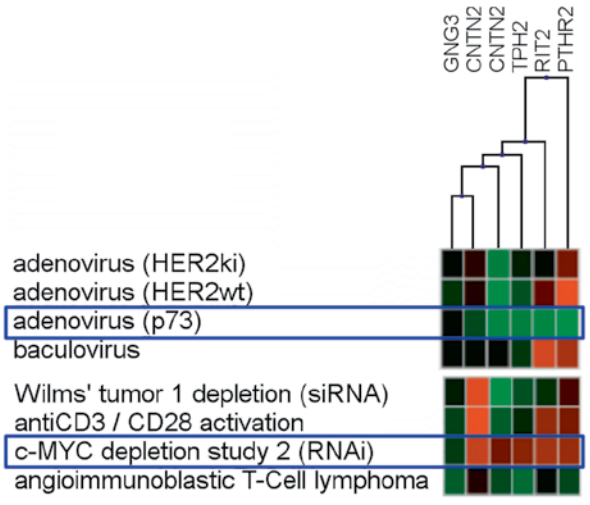

Figure 7. c-Myc and p73 regulate the neurosecretory PD markers. Fragment of an expression profile meta-analysis heat map illustrates the effect of c-Myc depletion and p73 overexpression on the expression of a subset of the PD differential markers. All affymetrix arrays with human gene sets were selected as input to Genevestigator. Red/green, elevated/diminished expression following treatment.

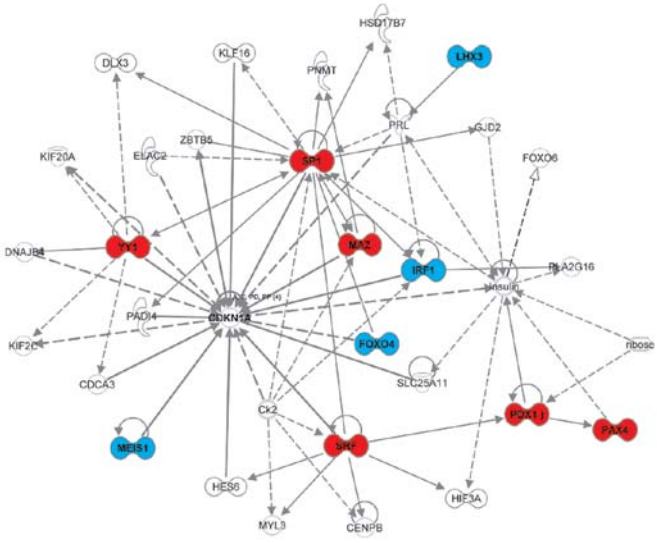

Figure 8. Highly regulated transcription factors have multiple functional connections with p21/CDKN1A. A network was constructed around the TFs which were most highly activated (red) or suppressed (blue) in PD tumours. White nodes were added automatically during the network construction process.

including large cell neuroendocrine carcinoma (LCNEC) and many small-cell lung cancers (SCLC), are well established (27). Indeed, neuroendocrine markers, such as chromogranin, are used to histologically classify LCNEC and SCLC (27). Our observation that certain neuroendocrine markers are elevated in recurrent NSCLC suggests that this system may need to be expanded to prevent the misclassification of NSCLC tumours. In support of this, the clinical aggressiveness of adenocarcinomas that contain a subpopulation of neuroendocrine cells has previously been suggested (28,29).

Our initial evidence for a connection between neuroendocrine markers and the progressive tumour phenotype came from the expression levels of specific neuroendocrine genes. Of the 23 genes which were robust in discriminating PD from DF expression profiles, 8 are involved in neuroendocrine phenomena. For example, TPH2 encodes a key enzyme in serotonin synthesis (30) and MTNR1A, a melatonin receptor gene, regulates circadian rhythm and hormone release (31). Potential biomarkers which have roles in more typical cancer-associated pathways (motility/apoptosis/growth factor signalling) were also identified in this screen, however, the neuroendocrine features of progressive tumours were further supported by our results on transcriptional control, discussed presently.

Transcriptional networks contributing to the progressive phenotype. Transcription factor activities (for normal, cancer, PD and DF samples) were inferred using TF-gene connectivities and the whole microarray dataset. Interestingly, many of the most selective TFs for the progressive phenotype have previously been implicated in neuroendocrine cell physiology and/or cell transformation/proliferation, despite not having focussed solely on the potential PD biomarkers in this analysis. For example, SP1 mediates the expression of the neuroendocrine marker chromogranin B (32), PDX1/IPF1 and PAX4 expression is controlled by the neuron-restrictive silencer element [which suppresses the expression of neural genes in most non-neural tissues; (33)] and the OCT1/2/3 transcription factors have roles throughout the neuroendocrine system (34). Notably, OCT2 mediates the differentiation of neuroendocrine brain cells. The activated TFs SOX9 and MAZ have elevated activity in prostate cancer (35,36), the latter playing a role in prostate neuroendocrine cancer and also the SP1-dependent regulation of parathyroid hormone-related peptide receptor expression (37). The potential function of the MAZ/SP1-PTHR pathway in the persistence of the progressive phenotype, in light of the expression of parathyroid hormone-like peptides by all major lung cancer cell types (38), is discussed later. Conversely, MEIS1 and IRF1, whose activity was down-regulated in progressive tumours, are known tumour suppressors (39,40), albeit through roles in leukaemia and the immune system, respectively. Finally, LEF1 is known to enhance the metastatic potential of lung adenocarcinoma cells (41). Since the activity of this WNT/TCF pathway component was strongly predictive for lung cancer, but did not discriminate PD from DF, it is anticipated that additional pro-metastatic genes/proteins may be required for recurrence in PD tumours.

Persistence networks of progressive cancer cells. There are a multitude of differences between any given pair of lung tumours, manifested at molecular, cellular and histological levels. Our analysis, however, shows that consistent changes (for example, in apoptotic or neuroendocrine genes) occur within PD tumours that distinguish them from matched DF counterparts. Convergent cellular mechanisms that allow the transformed phenotype to persist after surgery and which mediate the robust expression of our biomarker set in this heterogeneous background may therefore exist and potential mechanisms were identified though meta-analysis of published microarray studies and network analysis methods.

Genevestigator software demonstrated a negative correlation of Myc expression and p73 activation with 5 and 6 of our PD markers, respectively. Myc and p73 could, therefore, have an important role in regulating this cluster of genes. p73 isoforms have either pro- or antiapoptotic function (42), the p73 locus (1p36.33) is frequently deleted in squamous cell carcinoma (43) and induction of p73 has been shown to inhibit the proliferation of p53-mutant cells (44). Indeed, Myc has been shown to enhance p73 expression (45) and expression of dominant-negative p73 isoforms in patients with mutant-p53 ovarian cancer is associated with worse recurrence-free survival (46). Nonetheless, the possibility that the inhibition of a p73/Myc-based pathway could play a role in progressive lung cancer needs further verification.

Further support for a role of altered Myc signalling in PD tumours came from pathway analysis (Fig. 6). Many connections are observed from the PD-activated transcription factors to Myc and CDKN1A (which encodes p21; Fig. 8), suggesting Myc and CDKN1A as direct targets of these TFs in tumours that have a high probability of progression. Several studies have shown that lung tumours expressing low levels of p21 are associated with poor clinical outcome (15,47-49). p21 activation decreases cellular proliferation and decreases apoptosis in airway epithelial cells (50), although p21 may be proapoptotic under different cellular contexts (51,52). Sp1, which was one of the transcription factors most associated with PD tumours, is known to repress p21 expression in the presence of Myc and the histone deacetylase HDAC1 (52). Sp1 binds to its own promoter (see later), which may ensure its stable upregulation in progressive tumours.

The neuroendocrine markers (TPH2, RIT2, CNTN2, GPG3, PTH2R) exhibit close links with cAMP and Ca^{++} signalling. TPH2 is a rate-limiting enzyme in serotonin biosynthesis whose expression is itself regulated by cAMP in a manner dependent on the cancer-associated TF complex NF-Y/SP1 (30). Autocrine or paracrine signalling by serotonin or parathyroid hormone (via PTH2R) could sustain cellular cAMP levels and the expression of the neuroendocrine genes. Interestingly, cAMP, Ca^{++} and our PD-enriched TFs also interact with G2/M controlling proteins Rb, Cdc2 and Cyclin A.

Auto-regulatory TFs that were suggested by analysis of promoters and our transcription network include YY1, IRF1, SP1 and NF-Y (data not shown). The latter pair have previously been shown to form an autoregulatory transcription loop. MAZ, CREB, NF-Y and SP1 interact with histones and the histone acetylation/methylation machinery (Fig. 6). Epigenetic mechanisms and stable autoregulatory transcriptional loops may provide a further means for self-sustained growth of an established transformed phenotype after tumour resection. Since progressive tumours may develop at a distance from the original tumour, this internal rewiring of the epigenetic and transcriptional machinery may desensitise progressive tumour cells to the

altered environment encountered by mobile transformed lung cancer cells. This hypothesis, though intriguing, requires further testing *in vivo*.

Collectively, our systematic analyses provided a list of potential biomarkers for lung cancer recurrence or prognosis which require validation in a separate set of PD/DF samples. This identified 23 genes, a series of transcription factors (Fig. 3). When combined with known biological information, these genes/proteins provide a comprehensive understanding of changes in the regulatory processes underlying lung cancer. Several of the genes were involved in serotonin homeostasis and its regulatory pathways. Strong links between the identified biomarkers, transcriptional autoactivation and epigenetic processes were evident and may have implications for the understanding of cancer.

In conclusion, the analyses performed suggest that neuroendocrine signalling may have a role in the survival of aggressive lung cancer cells, and that the neuroendocrine phenotype may be connected to cancer-associated dysregulation of Myc and CDKN1A/p21.

Acknowledgements

The European Early Lung Cancer (EUELC) project was supported by a Framework V grant from the European Union (QLG1-CT-2002- 01735) and the Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 258677 – CURELUNG project. This research was also supported by the Roy Castle Lung Cancer Research Foundation. Z.D. was sponsored by the School of Computer Sciences, Universiti Sains Malaysia. We would like to thank the individuals who participated in this research and the clinicians and support staff who made this study possible.

Collaborators

The collaborating members of the European Early Lung Cancer (EUELC) Study Group are: Christian Brambilla (INSERM U823, Albert Bonniot Institute, Grenoble, France); Yves Martinet (Centeral Hospitalier Universitaire de Nancy, France); Frederik B. Thunnissen (Canisius Wilhelmina Ziekenhuis, Nijmegen, The Netherlands); Peter J. Snijders (University Hospital Vrije Universiteit, Amsterdam, The Netherlands); Gabriella Sozzi (Department of Experimental Oncology, Milan, Italy); Angela Risch (German Cancer Research Centre, Heidelberg, Germany); Heinrich D. Becker (German Cancer Research Centre, Heidelberg, Germany); J. Stuart Elborn (Belfast City Hospital, Belfast, United Kingdom); Luis M. Montuenga (University of Navarra, Pamplona, Spain); Ken J. O'Byrne (St. James Hospital, Dublin, Ireland); David J. Harrison (University of Edinburgh, Edinburgh, United Kingdom); Jacek Niklinski (Medical Academy of Bialystok, Bialystok, Poland); and John K. Field (Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, United Kingdom).

References

1. ONS: Mortality Statistics: Deaths registered in England and Wales (Series DR), 2008. 2009.
2. Tobias J and Hochhauser D: Cancer and its Management. John Wiley & Sons Ltd., 2005.
3. Field JK, Liloglou T, Niaz A, *et al*: EUELC project: a multicentre, multipurpose study to investigate early stage NSCLC, and to establish a biobank for ongoing collaboration. Eur Respir J 34: 1477-1486, 2009.
4. Maeda R, Yoshida J, Ishii G, Hishida T, Nishimura M and Nagai K: Prognostic impact of intratumoral vascular invasion in non-small cell lung cancer patients. Thorax 65: 1092-1098, 2010.
5. Harpole DH Jr, Herndon JE II, Young WG Jr, Wolfe WG and Sabiston DC Jr: Stage I nonsmall cell lung cancer. A multivariate analysis of treatment methods and patterns of recurrence. Cancer 76: 787-796, 1995.
6. Lu Y, Lemon W, Liu PY, *et al*: A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med 3: e467, 2006.
7. Guo NL, Wan YW, Tosun K, *et al*: Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. Clin Cancer Res 14: 8213-8220, 2008.
8. Shedden K, Taylor JM, Enkemann SA, *et al*: Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 14: 822-827, 2008.
9. Wilkerson MD, Yin X, Hoadley KA, *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.
10. Sun SS, Hsieh JF, Tsai SC, Ho YJ, Lee JK and Kao CH: Cytokeratin fragment 19 and squamous cell carcinoma antigen for early prediction of recurrence of squamous cell lung carcinoma. Am J Clin Oncol 23: 241-243, 2000.
11. Foa P, Fornier M, Miceli R, *et al*: Tumour markers CEA, NSE, SCC, TPA and CYFRA 21.1 in resectable non-small cell lung cancer. Anticancer Res 19: 3613-3618, 1999.
12. Bryant CM, Albertus DL, Kim S, *et al*: Clinically relevant characterization of lung adenocarcinoma subtypes based on cellular pathways: an international validation study. PLoS One 5: e11712, 2010.
13. Vinolas N, Molina R, Galan MC, *et al*: Tumor markers in response monitoring and prognosis of non-small cell lung cancer: preliminary report. Anticancer Res 18: 631-634, 1998.
14. Tamura M, Sawabata N, Kobayashi S, *et al*: Prognostic significance of p21 protein expression in patients with pulmonary squamous cell carcinoma following induction chemotherapy. Ann Thorac Cardiovasc Surg 13: 9-14, 2007.
15. Wu DW, Liu WS, Wang J, Chen CY, Cheng YW and Lee H: Reduced p21(WAF1/CIP1) via alteration of p53-DDX3 pathway is associated with poor relapse-free survival in early-stage human papillomavirus-associated lung cancer. Clin Cancer Res 17: 1895-1905, 2011.
16. Garcia-Yuste M, Matilla JM and Gonzalez-Aragoneses F: Neuroendocrine tumors of the lung. Curr Opin Oncol 20: 148-154, 2008.
17. Rekhtman N: Neuroendocrine tumors of the lung: an update. Arch Pathol Lab Med 134: 1628-1638, 2010.
18. Berger J, Hautaniemi S, Jarvinen A-K, Edgren H, Mitra S and Astola J: Optimized LOWESS normalization parameter selection for DNA microarray data. BMC Bioinformatics 5: 194, 2004.
19. Hancock D, Wilson M, Velarde G, *et al*: maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination. BMC Bioinformatics 6: 264, 2005.
20. Celton M, Malpertuy A, Lelandais G and de Brevern A: Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics 11: 15, 2010.
21. McMaster A, Jangani M, Sommer P, *et al*: Ultradian cortisol pulsatility encodes a distinct, biologically important signal. PLoS One 6: e15766, 2011.
22. Sanguinetti G, Lawrence ND and Rattray M: Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. Bioinformatics 22: 2755-2781, 2006.
23. IPA (Ingenuity® Systems, www.ingenuity.com).
24. Siu DC and Laurance M: Rapid generations of de novo biological pathways from large-scale gene expression data using the Ingenuity Pathways Analysis application. Proc Am Assoc Cancer Res 45: 756-b-, 2004.
25. Hruz T, Laule O, Szabo G, *et al*: Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics: 420747, 2008.
26. Van Lommel A, Bolle T, Fannes W and Lauweryns JM: The pulmonary neuroendocrine system: the past decade. Arch Histol Cytol 62: 1-16, 1999.
27. Travis WD: Advances in neuroendocrine lung tumors. Ann Oncol 21 (Suppl. 7): vii65-71, 2010.

28. Pelosi G, Pasini F, Sonzogni A, *et al*: Prognostic implications of neuroendocrine differentiation and hormone production in patients with Stage I nonsmall cell lung carcinoma. Cancer 97: 2487-2497, 2003.
29. 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 USA 98: 13790-13795, 2001.
30. Cote F, Schussler N, Boularand S, *et al*: Involvement of NF-Y and Sp1 in basal and cAMP-stimulated transcriptional activation of the tryptophan hydroxylase (TPH) gene in the pineal gland. J Neurochem 81: 673-685, 2002.
31. Johnston JD, Messager S, Barrett P and Hazlerigg DG: Melatonin action in the pituitary: neuroendocrine synchronizer and developmental modulator? J Neuroendocrinol 15: 405-408, 2003.
32. Mahapatra NR, Mahata M, Ghosh S, Gayen JR, O'Connor DT and Mahata SK: Molecular basis of neuroendocrine cell type-specific expression of the chromogranin B gene: crucial role of the transcription factors CREB, AP-2, Egr-1 and Sp1. J Neurochem 99: 119-133, 2006.
33. Kemp DM, Lin JC and Habener JF: Regulation of Pax4 paired homeodomain gene by neuron-restrictive silencer factor. J Biol Chem 278: 35057-35062, 2003.
34. Andersen B and Rosenfeld MG: POU domain factors in the neuroendocrine system: lessons from developmental biology provide insights into human disease. Endocr Rev 22: 2-35, 2001.
35. Wang H, McKnight NC, Zhang T, Lu ML, Balk SP and Yuan X: SOX9 is expressed in normal prostate basal cells and regulates androgen receptor expression in prostate cancer cells. Cancer Res 67: 528-536, 2007.
36. Hu Y, Wang T, Stormo GD and Gordon JI: RNA interference of achaete-scute homolog 1 in mouse prostate neuroendocrine cells reveals its gene targets and DNA binding sites. Proc Natl Acad Sci USA 101: 5559-5564, 2004.
37. Williams L and Abou-Samra A: The transcription factors SP1 and MAZ regulate expression of the parathyroid hormone/parathyroid hormone-related peptide receptor gene. J Mol Endocrinol 25: 309-319, 2000.
38. Brandt DW, Burton DW, Gazdar AF, Oie HE and Deftos LJ: All major lung cancer cell types produce parathyroid hormone-like protein: heterogeneity assessed by high performance liquid chromatography. Endocrinology 129: 2466-2470, 1991.
39. Lasa A, Carnicer MJ, Aventín A, *et al*: MEIS 1 expression is downregulated through promoter hypermethylation in AML1-ETO acute myeloid leukemias. Leukemia 18: 1231-1237, 2004.
40. Stang MT, Armstrong M, Liu Y, Yan P and Yim JH: IRF-1 suppresses tumor formation and protects against tumor rechallenge in a murine model of breast carcinoma: the role of tumor specific immunity. J Surg Res 121: 300, 2004.
41. Nguyen DX, Chiang AC, Zhang XH-F, *et al*: WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis. Cell 138: 51-62, 2009.
42. Rufini A, Agostini M, Grespi F, *et al*: p73 in cancer. Genes Cancer 2: 491-502, 2011.
43. Nomoto S, Haruki N, Kondo M, Konishi H and Takahashi T: Search for mutations and examination of allelic expression imbalance of the p73 gene at 1p36.33 in human lung cancers. Cancer Res 58: 1380-1383, 1998.
44. Irwin M, Marin MC, Phillips AC, *et al*: Role for the p53 homologue p73 in E2F-1-induced apoptosis. Nature 407: 645-648, 2000.
45. Zaika A, Irwin M, Sansome C and Moll UM: Oncogenes induce and activate endogenous p73 protein. J Biol Chem 276: 11310-11316, 2001.
46. Concin N, Hofstetter G, Berger A, *et al*: Clinical relevance of dominant-negative p73 isoforms for responsiveness to chemotherapy and survival in ovarian cancer: evidence for a crucial p53-p73 crosstalk in vivo. Clin Cancer Res 11: 8372-8383, 2005.
47. Caputi M, Esposito V, Baldi A, *et al*: $p21^{waf1/cip1mda-6}$ expression in non-small-cell lung cancer: relationship to survival. Am J Respir Cell Mol Biol 18: 213-217, 1998.
48. Shoji T, Tanaka F, Takata T, *et al*: Clinical significance of p21 expression in non-small-cell lung cancer. J Clin Oncol 20: 3865-3871, 2002.
49. Komiya T, Hosono Y, Hirashima T, *et al*: p21 expression as a predictor for favorable prognosis in squamous cell carcinoma of the lung. Clin Cancer Res 3: 1831-1835, 1997.
50. Blundell R, Harrison DJ and O'Dea S: p21(Waf1/Cip1) regulates proliferation and apoptosis in airway epithelial cells and alternative forms have altered binding activities. Exp Lung Res 30: 447-464, 2004.
51. Liu S, Bishop WR and Liu M: Differential effects of cell cycle regulatory protein p21(WAF1/Cip1) on apoptosis and sensitivity to cancer chemotherapy. Drug Resist Updat 6: 183-195, 2003.
52. Ocker M and Schneider-Stock R: Histone deacetylase inhibitors: signalling towards $p21^{cip1/waf1}$. Int J Biochem Cell Biol 39: 1367-1374, 2007.