Identification of alternative splicing alterations in small cell lung cancer

Marina Reixachs i Solé

Treball de Fi de Grau

Biologia Humana

Universitat Pompeu Fabra

Supervisor: Eduardo Eyras

Regulatory Genomics Group

Research Programme on Biomedical Informatics, GRIB (IMIM-UPF)

Abstract

Lung cancers cause 1,5 million casualties per year worldwide. Despite their heterogeneity, lung cancers are classified into two classes as small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which include lung adenocarcinoma (LUAD) and lung squamous cell cancer (LUSC). SCLC is the most aggressive type of lung cancer reaching average survival rates of 5% after 5 years from the time of diagnosis. The lack of knowledge about the underlying tumorigenic mechanisms and the lack of effective treatments make the situation more dramatic for SCLC tumours. Our main goal is to obtain a specific splicing signature for SCLC that may provide novel molecular targets for prognosis and therapy. For this purpose, we used the iso-kTSP algorithm, a recently developed computational method able to identify transcript isoform changes across different samples. This comparison-based approach allows to classify RNA-seq data from different tumour or normal samples by establishing a decision rule based on the relative ordering in a ranking of isoform expression values. We applied this method to samples from different lung cancers: SCLC, LUAD and LUSC; and also to normal lung samples. Our results revealed a set of distinct alternative splicing patterns in SCLC with potential functional relevance. This work shows that identification of alterations in alternative splicing can shed light on the study of new molecular mechanisms to develop prognostic and therapeutic targets of SCLC tumours.

Introduction

Lung cancer is the most commonly diagnosed type of cancer worldwide and results in the largest number of cancer related deaths with 1,5 million casualties per year (1,2). Despite their heterogeneity, lung cancers are classified in two major groups which are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). This classification is based on their different cellular origin, metastatic spread and therapeutic response (3,4). Non-small cell carcinomas account for more than 85% of diagnosed cases showing two predominant histological entities, lung adenocarcinoma (LUAD) and squamous cell cancer (LUSC), both originated from glandular epithelial cells (5,6). Moreover, adenocarcinoma is the most common type of lung cancer among non-smokers. Even though small cell lung carcinomas only represent less than 15% of lung cancers, it is the most aggressive type of lung cancer reaching average survival rates of 5% after 5 years from the time of diagnosis. SCLC is thought to have a neuroendocrine origin since many SCLC tumours express neuroendocrine markers (3).

In the last 10 years a lot of work has been done in the identification of molecular pathways for the development of these cancers. For LUAD tumours, somatic mutations that drive oncogenesis have been described for EGFR, KRAS, ALK, RET and ROS (7–9). Some of these genes such as EGFR and ALK are in fact being therapeutically targeted with satisfactory clinical outcomes (10,11). However, for a considerable amount of lung adenocarcinomas there is a lack of knowledge about their underlying tumorigenic mechanisms or their targeted therapeutic strategies have not been successful. Tumours that do not present any driver mutation described so far are called “pan-negative” and patients presenting this kind of tumours cannot benefit from any targeted therapy. In SCLC the lack of known actionable mechanisms is even more dramatic since the vast majority of SCLC tumours are pan-negative. Analysis of multiple SCLC samples has been crucial for the identification of alterations in tumour suppressors such as TP53 and RB or the MYC gene family (12,13). Unfortunately, the mentioned targets are not currently clinically actionable.

The relevance of the actual transcript repertoire in cancer is clear since it is a major modulator of the functional content of tumour cells. Previous studies have revealed a frequent fusion in transcripts of the RET gene in pan-negative LUAD samples that can be targeted with known tyrosine kinase inhibitors (14,15). This shows that identification of alterations at the transcript level can provide new insights into targeted therapeutic strategies, especially in pan-negative tumours.

Changes in the transcriptome are often reflected as alternative splicing abnormalities in tumours. Alternative splicing both involves *cis*-regulatory sequences and interactions with *trans*-acting factors, known as 'splice factors' (16). Accordingly, splicing alterations can be classified as *cis* and *trans* depending on whether they happen at sequence or regulatory machinery level, respectively.

Alterations in alternative splicing are emerging as tools for understanding cancer biology (16,17). It has been described that this post-translational process has a role in promoting angiogenesis (18), proliferation (19), cell invasion (20) and apoptosis avoidance (21), giving tumours a selective advantage. Moreover, aberrant splicing patterns can also be the source of therapeutic resistance as it is the case for RAF inhibitors resistance in melanomas expressing BRAF(V600E) splice variants (22).

Hypothesis

There is plenty of evidence of alternative splicing alterations that may have a role in SCLC tumorigenic process. Splicing factors SRSF1 and hRNPA1 are target genes of MYC transcription factor, which appears amplified in SCLC tumours (23). Consequently, these splicing factors may experiment expression changes leading to multiple alternative splicing events with oncogenic effects (24). Similarly, mutations in RB may also result in splicing alterations, as the RB pathway is known to control RNA processing and splicing (25). In addition, a splicing variant of the cell adhesion molecule CADM1 has been found specifically expressed in SCLC samples, suggesting that CADM1 may promote malignant growth in SCLC tumours (26). A truncated splice variant of the silencing transcription factor NRSF has been also reported to appear specifically in SCLC tumours compared to other lung tumours (27), indicating an unpaired regulation of genes encoding adhesion molecules or neuroendocrine peptides expressed in SCLC (28).

Thus, we hypothesize that detection of splicing alterations will be relevant to explain oncogenic mechanisms in SCLC and it can further contribute to improving prognostic and therapeutic strategies.

Objectives and problem approach

Sometimes alternative splicing may involve complex patterns that cannot be explained by simple splicing events. As it has been shown for TP53 (29), alternative splicing alterations can be described in terms of isoform expression changes between tumour and normal samples. Additionally, recent work has revealed the importance of recurrent changes in the relative abundance of the transcript isoforms of a gene to provide novel signatures in cancer (30).

Our main goal is to obtain specific splicing patterns for SCLC through the analysis of isoform expression data from multiple samples of SCLC and other lung tumours. Furthermore, we aim that our results will help to elucidate some of the specific mechanisms that lead SCLC tumorigenic process and will benefit the development of prognostic and therapeutic targets for these tumours.

For this purpose, we used a computational approach consisting in a comparison-based algorithm which appeared to be robust for the identification of transcript isoform changes across multiple samples from different tumour types (30). Reversals in the relative abundance of alternative transcripts results in alterations of their relative order in a ranking of transcript expression (30). Accordingly, we identified switches in the relative expression ordering of different isoforms that allowed us to report several genes which appear altered at the transcript level in SCLC samples. Finally, we also performed function enrichment analysis with the resulting candidate genes in order to explore the potential role of the identified splice variants in the oncogenic process. A schematic representation of the approach is shown in (Figure 1).

Materials and methods

Data collection

Data from samples of three different lung cancer types are used: lung adenocarcinoma (LUAD and LADC), lung squamous cell carcinoma (LUSC) and small cell lung carcinoma (SCLC) (Table 1). For LUAD and LUSC, data for both tumour and normal samples were downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). In this context, normal samples are those obtained from LUAD or LUSC patients that do not present cancerous characteristics. Data from SCLC tumour samples and cell lines were provided by (31) together with LADC samples (15).

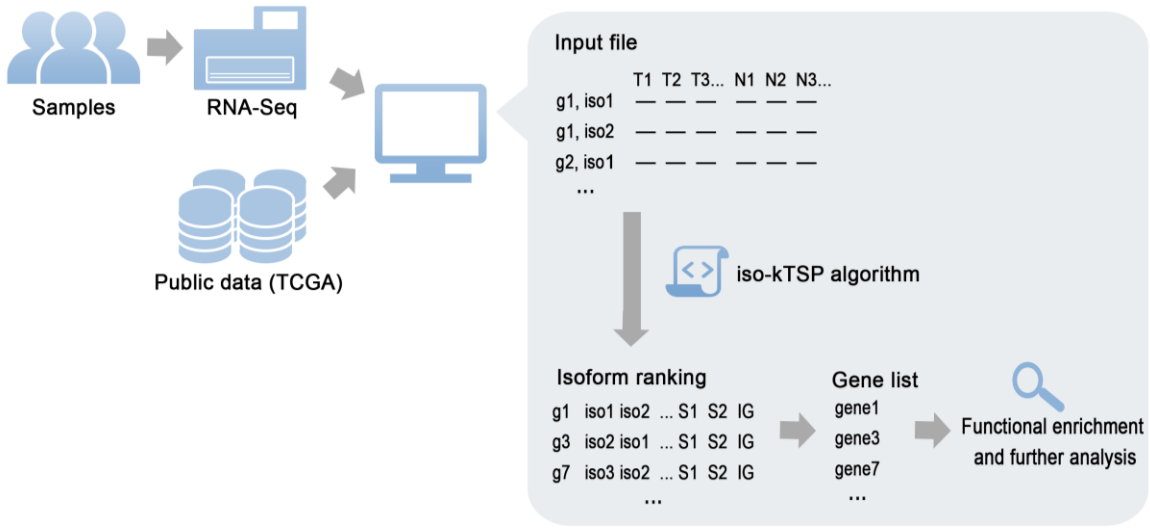

Figure 1. Schematic representation of the process. Data from public fonts and RNA-seq data from samples of lung cancer patients were analysed using a computational based approach, the iso-kTSP algorithm, in order to obtain isoform expression rankings that have been further analysed to identify recurrent isoform switches and possible functional implications in SCLC cancer.

Samples		Num.	Source	Ref
LUAD	tumour	534	TCGA	(7)
	normal	58	TCGA	(7)
LADC	tumour	169	Kohno et al.	(15)
LUSC	tumour	504	TCGA	(32)
	normal	51	TCGA	(32)
SCLC	fresh tumour	20	Iwakawa et al.	(31)
	cell lines	23	Iwakawa et al.	(31)

Table 1. Summary of all data available with their corresponding references.

The abundance of every transcript in the samples is expressed in terms of transcripts per million (TPM). For an isoform i, the TPM value can be obtained by multiplying by 10^6 the fraction of transcripts τ_i as described in (33):

$$\tau_i = \frac{\nu_i}{\ell_i} \cdot \left(\sum_j \frac{\nu_j}{\ell_j} \right)^{-1}$$

Where ν_i is the fraction of nucleotides and ℓ_i is the transcript length in nucleotides.

The iso-kTSP algorithm

In order to detect changes in alternative splicing that cause reversals in the relative expression of the transcript in different cancer samples, the iso-kTSP algorithm developed by Sebestyén et al. (30) was used (software and additional information available at https://bitbucket.org/regulatorygenomicsupf/iso-ktsp).

The iso-kTSP algorithm is based on the Top-Scoring Pairs (TSP) method that was originally developed to measure consistent gene expression reversals (34,35). It is a comparison-based approach that allows sample classification based on a decision rule. In the present work, the rule is determined by the relative ordering of isoform expression values. To adapt the method to RNA-seq data, in which there are many transcripts with zero reads, scoring calculations described below are different from the ones in (34,35).

The algorithm stores the isoform expression ranking according to TPM values of each sample from two possible classes,C_m, $m = 1,2$. Then, for each isoform-pair $I_{g,i}, I_{g,j}$ in each gene *g*, iso-kTSP extracts the frequency of the two possible relative orders and calculates the following score:

$$S_1(I_{g,i}, I_{g,j}) = P(I_{g,i} > I_{g,j}|C_1) + P(I_{g,i} < I_{g,j}|C_2) - 1$$

Where $P(I_{g,i} > I_{g,j}|C_1)$ is the frequency in which $I_{g,i}$ appears later than $I_{g,j}$ in the expression ranking of C_1 and $P(I_{g,i} < I_{g,j}|C_2)$ is the frequency in which $I_{g,i}$ appears earlier than $I_{g,j}$ in C_2. This score S_1 estimates how probable is for a pair of isoforms from the same gene to change their relative order in the expression ranking between the two classes (C_1 and C_2).

Additionally, a second score S_2 is calculated. The position in the ranking of an isoform in sample S_a belonging to a class C_m is defined as $R(I_{g,i}|S_a, C_m)$. Then, the average rank difference between two isoforms per class is:

$$g(I_{g,i}, I_{g,j}|C_m) = \frac{1}{|C_m|}\sum_a \left(R(I_{g,i}|S_a, C_m) - R(I_{g,j}|S_a, C_m)\right)$$

When $|C_m|$ is the number of samples in class C_m. Thus, the final S_2 score is the difference between the average rank difference of two isoforms in each class (C_1, C_2), as defined previously (35).

$$S_2(I_{g,i}, I_{g,j}) = |g(I_{g,i}, I_{g,j}|C_1) - g(I_{g,i}, I_{g,j}|C_2)|$$

This second score provides an estimate for the magnitude of the isoform expression reversal for isoform pairs from the same gene between two classes.

At this point, all pairs of isoforms are sorted according to score S_1 and in the case of a tie, the second score S_2 is used. Only pairs of isoform from the same gene are considered for sorting and a single isoform pair per gene is selected for the global isoform expression ranking, so every gene can only be listed once.

Each one of these isoform pairs provides a rule for the classification subject to their relative order. The semantics for a rule is that if the first isoform has a lower expression than the second, it is predicted as C_1 and otherwise the prediction is C_2. Hence:

rule: $I_{g,j} < I_{g,i}$
$I_{g,j} < I_{g,i} \Rightarrow C_1$
otherwise $\Rightarrow C_2$

C_1 is the first class in the input file and generally accounts for SCLC samples, either fresh tumours or cell lines, and C_2 for other samples.

For instance, if C_1 refers to fresh tumour SCLC samples and C_2 to LADC samples, the prediction would be:

rule: $I_{g,j} < I_{g,i} \Rightarrow$ SCLC
otherwise $\Rightarrow$ LADC

In this case, the first isoform $I_{g,i}$ would be predicted as “SCLC” and the second one $I_{g,j}$ would be predicted as “LADC”.

In order to classify a new sample, each isoform-pair rule is evaluated against the isoform expression ranking of the new sample. Considering *k* isoform-pair rules, the classifier determines the class for which the data fulfills each rule. Final classification is established by majority voting. Predictive models always use an odd number of rules in order to avoid ties in the final classification. Following the proposed example, for *k = 3*:

rule 1: SCLC
rule 2: SCLC
rule 3: LADC
} classification: SCLC

The optimal number of *k* isoform pairs is obtained by performing a cross-validation analysis. In the cross-validation samples are split into training and testing sets containing the same number of samples from each class. The number of iterations in the cross-validation can be defined as *n* and indicates in how many groups the sample will be split. The algorithm is set to perform a 10-fold cross-validation (*n = 10*) by default which means that samples are partitioned in 10 equal size groups. Additionally, we performed a leave one-out (LOO) cross validation which is the same as using an *n* value equal to the amount of samples. Training sets contain samples from n-1 partitions and the remaining sample group forms the test set. At each iteration step, the top k-pairs ($k=1…k_{max}=50$, with *k* odd) of the isoform-pair ranking obtained from the training set is applied to the test set. In the case of LOO cross-validation, the model is obtained from all samples except one that is used for testing. The resulting predictions are compared with the actual labels of the samples, classes C_1 and C_2 in our case. Hence, for each class we have correctly and incorrectly predicted samples, that will be treated as true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The accuracy for each k is then:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Additionally, an information gain (IG) value is reported. This value reports an estimate of the predictive power of each isoform pair. It is calculated in terms of entropy reduction according to the samples that are correctly and incorrectly classified.

Finally, the global ranking is obtained with average values for all the mentioned parameters (S_1, S_2, IG and accuracy) obtained at each cross-validation step. The number of isoforms in the global ranking (*s)* was established to be 200. The final predictive model is extracted from the global ranking selecting the minimum odd number of isoform pairs, k_{opt}, with the highest average performance.

Input data and pre-processing

Input files contains a header with *m* labels (one label per sample) followed by the corresponding suffix that allows the program to distinguish between classes. The following rows have *m* + 1 columns, since the first column of each row shows the gene and isoform ids. Subsequent columns contain expression TPM values of the samples with the same order as in header (Figure 1).

As data is provided separately, a Perl script was applied to join data from two different classes. The script also enables to group isoforms from the same gene together, add class suffixes to the labels and select subsets of a specified number of random samples.

SCLC data from fresh tumours and cell lines was analysed separately against the remaining data sets. Additionally, three data sets with randomly merged adenocarcinoma, tumour and normal samples were created and analysed against both SCLC data (Figure 2).

We analysed balanced groups containing an equal number of samples from the two classes. The amount of samples was determined by the maximum number of samples in the less represented class. Hence, IG values are normalised and easier to interpret. To avoid bias due to sample selection, we obtained three randomized balanced groups for each combination of data sets 1 and 2.

Output files and information extraction

The output shows different lines that report the top scoring isoform-pairs for each iteration within the cross validation and the corresponding S_1, S_2, IG and detailed accuracy values. Next, the global isoform ranking with the average performance obtained in the cross-validation. The latter lines contain the minimal predictive model.

From each output file, we extracted the ranking obtained in the final single pair performance. As the scores of the isoform pairs on the lowest positions were high enough, all 200 pairs were considered. Only the genes that resulted from all of the three randomized input files, named r1, r2 and r3, are selected for further analysis, generating a unique isoform-pair list for every comparison between different data sets (Figure 2). For each gene list, the average of S_1 score values resulting from the three randomized input files was obtained as:

$$\overline{S_1} = \frac{S_1^{r1} + S_1^{r2} + S_1^{r3}}{3}$$

Finally, these unique lists were compared against each other to extract the desired information.

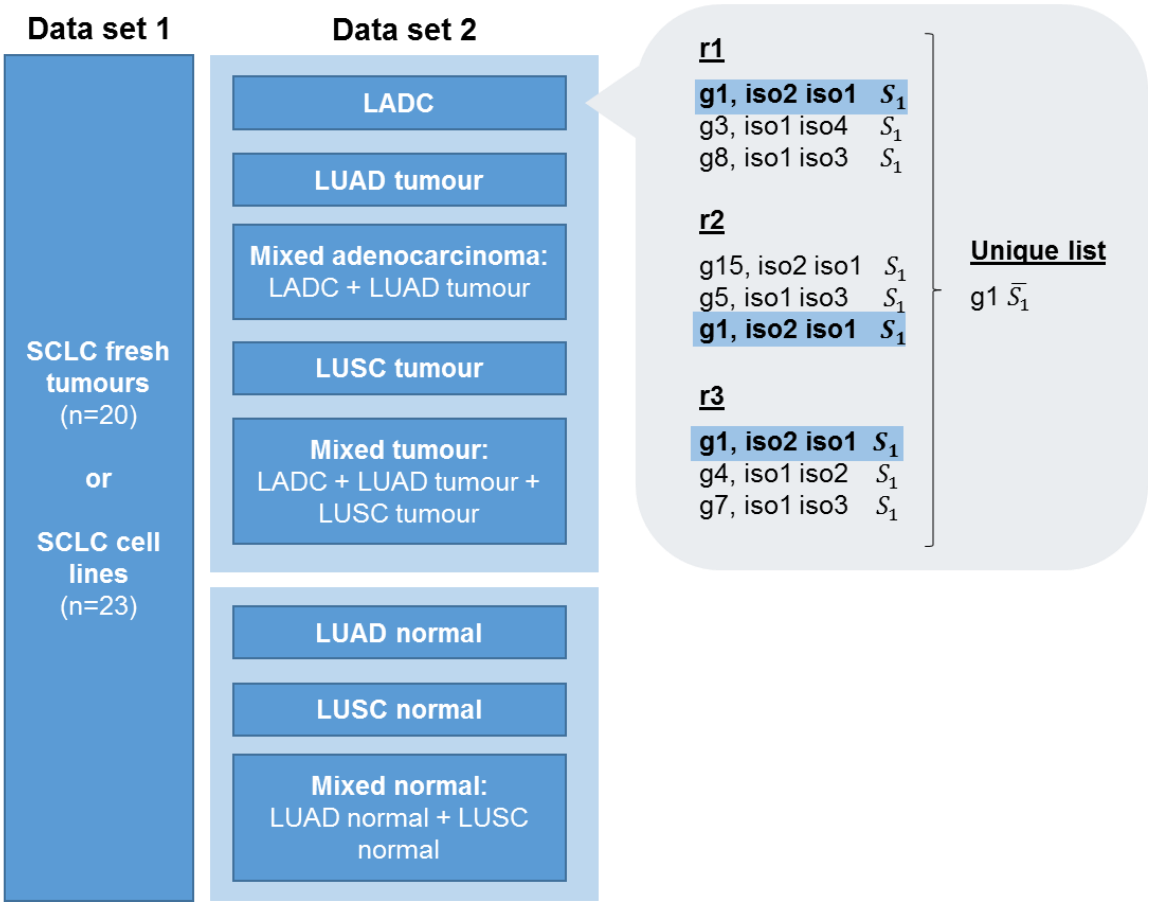

Figure 2. Samples from SCLC fresh tumours (n=20) forming the data set 1 were compared with an equal number of samples from the eight different data sets: LADC, LUAD tumour, mixed adenocarcinoma, LUSC tumour, mixed tumour, LUAD normal, LUSC normal and mixed normal. For each combination, three different randomized files (r1, r2 and r3) were analysed. The resulting rankings were put together in order to extract common isoforms. The scores of these unique lists were calculated as the average of the three scores from r1, r2 and r3 rankings: $\overline{S_1} = (S_1^{r1} + S_1^{r2} + S_1^{r3}) / 3$.

Enrichment analysis

The functional enrichment analysis was performed using the web-based application GOrilla (http://cbl-gorilla.cs.technion.ac.il). This online resource identifies enriched Gene Ontology (GO) terms in ranked gene lists and provides a p-value for each term. GO consists of three hierarchically structured vocabularies (ontologies) that describe gene products according to their related biological processes, cellular components and molecular functions (36).

In our case, we provided the listed candidate genes and a background consisting of all the genes from our data with more than one isoform. That is, the background is defined as all genes that could have appeared as an isoform switch. Additionally, we only considered GO terms with p-values smaller than 10^{-3}.

Results

SCLC presents recurrent alternative splicing isoform switches with high S_1 score and IG values

The iso-kTSP algorithm was applied to each one of the three different randomized files of different combinations of data sets 1 and 2 (Figure 2). The resulting ranked isoforms in all analyses performed showed S_1 score values greater than 0,85 and IG values greater than 0,6.

The rankings obtained from the three randomizations of each different joined data set were put together in a single list containing only the isoforms present in all the three resulting rankings. The resulting lists contain more than 100 isoforms out of the 200-isofom rankings which are common in the three outputs (Table 2). These results indicate that even though we selected random samples for the second data set, a considerable number of genes appear recurrently in all of the three analyses performed.

		Data set 1 (SCLC)	
		FT	CL
Data set 2	LADC	113	124
	LUAD tumour	130	105
	Mixed adenocarcinoma	128	106
	LUSC tumour	132	129
	Mixed tumour	122	112
	LUAD normal	142	159
	LUSC normal	147	165
	Mixed normal	149	153

Table 2. Summary of the total common genes in isoform rankings from the three randomizations (r1, r2 and r3) of different combinations of data sets 1 and 2.

For all the genes resulting from the intersection of the three randomized data sets, an average score was calculated. Regarding the average isoform-pair scores, it appears that S_1 score from the analysis of both SCLC fresh tumour and cell lines against normal samples distribute around values between 0,95-1. In contrast, analyses against tumour samples show a range of S_1 values from 0,85-1 (Figure 3). Consequently, we can infer that changes in relative isoform abundance are more consistent between SCLC and normal samples than with respect to tumour samples.

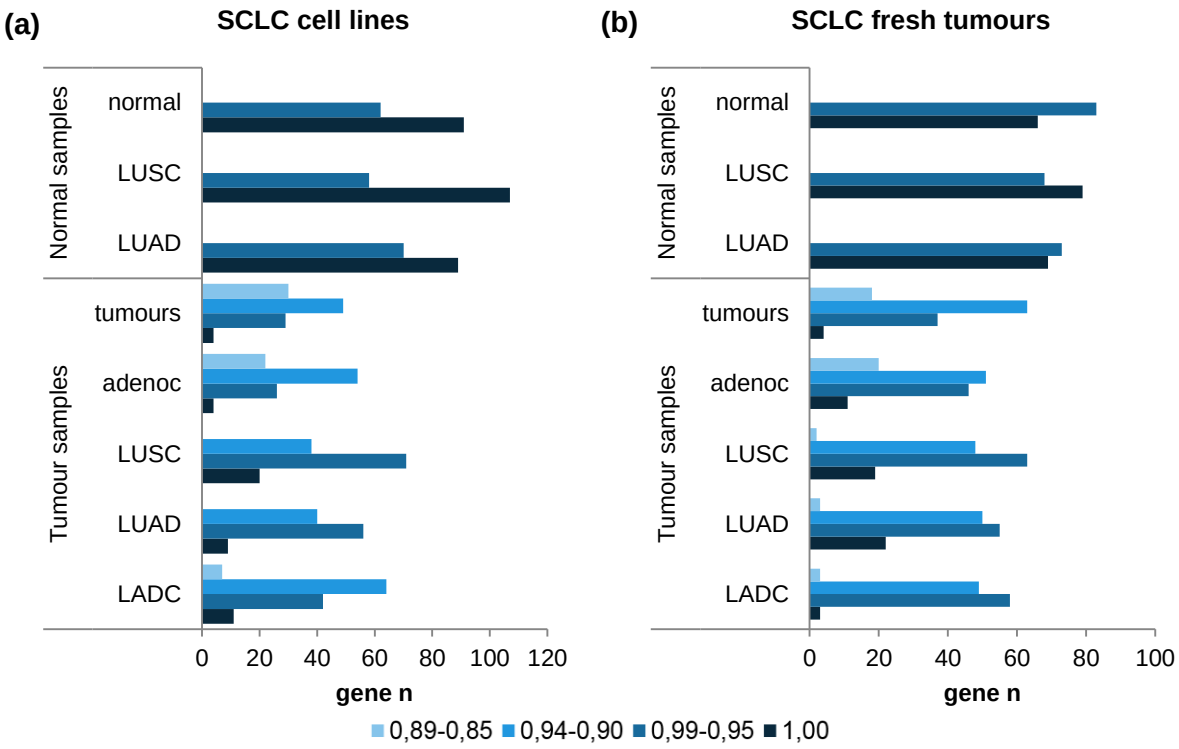

Figure 3. Representation of the number of genes presenting isoform switches with an average value within the indicated ranges, for every comparison between SCLC cell lines (a) or fresh tumours (b) against different data sets (LADC, LUAD tumour, mixed adenocarcinoma, LUSC tumour, mixed tumour, LUAD normal, LUSC normal and mixed normal) as described in (Figure 2).

In order to delimit the values of S_1 score and IG resulting from analyses between similar samples, we run the algorithm with sets of normal samples, from LUAD and LUSC, which should not present differences. All the obtained S_1 values are comprised within the range from 0,3 to 0,6 and IG values are smaller than 0,35. We used these values as further confirmation that IG and score values from the analysis of SCLC against other samples were significantly high.

Recurrent isoform changes are consistent among analysis of different data sets

Analysing the different types of SCLC samples independently, we found 324 unique genes resulting from all analyses with fresh tumour samples and 316 for cell lines. Since 130 of those genes are shared by the two SCLC sample classes, our analysis allowed us to identify a global total of 510 unique isoform switches.

For detection of the most recurrent isoform changes, we extracted the genes that appeared as a result of at least 7 out of the total 8 performed analysis for fresh tumours and cell lines. We obtained a list of 49 genes for fresh tumours and 46 genes for cell lines (see Supplementary table S1).

Additionally, we considered the results from analysis against tumour (LADC, LUAD, LUSC, mixed adenocarcinomas and mixed tumours data sets) and normal samples

(LUAD, LUSC and mixed normal data sets) separately. For fresh tumours we obtained 113 common genes for the normal group and 76 for the tumour group. Cell lines showed similar results, with 118 genes resulting from comparisons against normal samples and 68 from tumour (Table 3). We observe that there are more common genes that present isoform changes when comparing with normal samples rather than tumour samples. This can be interpreted as higher consistency among comparisons with normal samples.

SCLC cell lines samples present recurrent isoform changes that are also observed in fresh tumours

In order to find isoform changes in samples from SCLC fresh tumours that could be also identified in cell lines, we compared the previously mentioned common gene lists from the analysis against tumour and normal data sets. In the case of tumour samples, 14 genes were shared by both fresh tumour and cell line samples. For analysis between SCLC and normal samples, we obtained a list of 37 genes present in both cell lines and fresh tumour analysis (Figure 4). In addition, we can observe that 6 genes are present in both tumour and normal lists, namely, ASPH, DICER1, MARCH8, PPHLN1, RRAS2 and UHMK1.

Data set 1 (SCLC)	Data set 2	Group name	Number of common genes
Fresh tumours	LADC LUAD tumour LUSC tumour Mixed adenocarcinoma Mixed tumour	Tumour samples	76
	LUAD normal LUSC normal Mixed normal	Normal samples	113
Cell lines	LADC LUAD tumour LUSC tumour Mixed adenocarcinoma Mixed tumour	Tumour samples	68
	LUAD normal LUSC normal Mixed normal	Normal samples	118

Table 3. Number of common genes resulting from comparing all analyses of SCLC fresh tumours or cell lines against tumour or normal samples. Tumor samples include isoform rankings from LADC, LUAD tumour, mixed adenocarcinoma and mixed tumour. The group named samples is formed by rankings from LUAD normal, LUSC normal and mixed normal results.

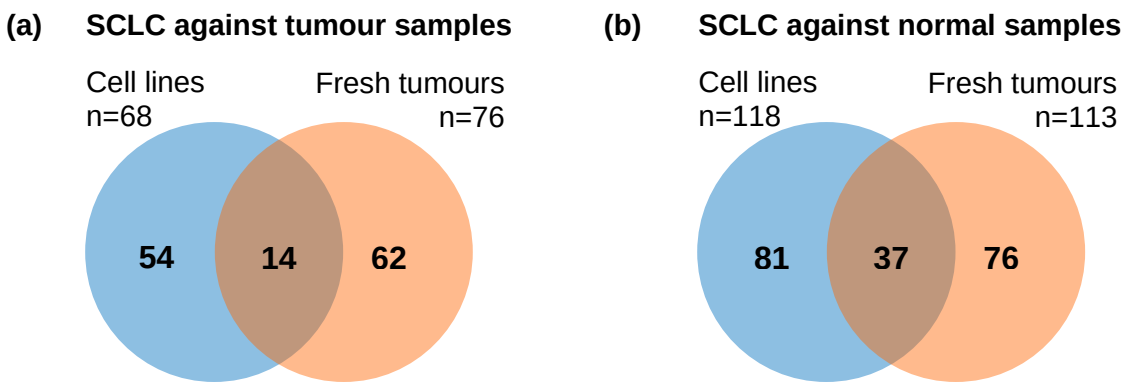

Figure 4. Diagrams showing common genes from the lists obtained with SCLC cell lines and fresh tumours against tumour (a) or normal samples (b). Complete gene lists of the intersections can be found at (Supplementary table S2).

Several processes appear enriched in both fresh tumour and cell line SCLC samples

Common gene lists from either SCLC fresh tumour or cell lines samples against normal or tumour samples as defined before in (Table 3) were used for the functional enrichment analysis together with the corresponding background. We obtained different processes that appeared enriched in the four analyses from the different data sets (for more details see Supplementary tables S3-6). Note that due to the hierarchical organization of GO term classification some of the processes appear redundantly as they belong to lower hierarchical levels.

Discussion

With this work, we identified 510 unique genes that show differential transcript expression in SCLC compared to other lung cancers or normal tissue. Moreover, results from different analyses performed seem to be consistent and present high score and IG values. The obtained results indicate that our approach was successful for identification of isoform switches using RNA-seq data from SCLC. Studies like the one presented here have a substantial importance, especially for SCLC, since much of the knowledge about its molecular mechanisms was unveiled as high throughput techniques emerged. Consequently, the development and application of new computational methods is also crucial to elicit relevant information from the growing available data.

Referring to enrichment analysis, many of the enriched processes identified were related to regulation of metabolic processes. More precisely, a negative regulation of catabolic processes was found for fresh tumours against normal samples. Interestingly, in all

analyses neuronal-related processes were enriched, such as neuron projection, axonal guidance or neural progenitor cells division linked to genes FGFR2 and DIXDC1. In addition, fresh tumours showed enriched processes ligated to RNA, like the regulation of its stability in relation to QKI, DICER1, SYNCRIP, BRF1 and PAIP1 genes. Moreover, a process of positive regulation of RNA polymerase II in response to hypoxia was described. We obtained other enriched processes such as angiogenesis, regulation of phosphoprotein phosphatase activity or cell signaling regulation. Probably, some of these processes might have a role in giving selective advantage to tumours and helping tumour progression. In cell lines, we also found that aberrant splicing occurred in genes related with lung differentiation processes that are not identified in fresh tumour samples. Some of them are associated with NUMA1, NFIB, RBPJ and NUMB genes.

Although a lot of work has to be done to directly link the identified genes and processes with SCLC oncogenic mechanisms, in some cases their relation with cancer can be defined. For instance, unpaired stabilization of mRNA has been described as a mechanism that triggers invasiveness in breast cancers (37). Another remarkable example is the case of VEGFA gene, which appeared in association with processes such as angiogenesis and axonal guidance, both related with tumourigenesis and cancer progression (38,39). Additionally, alternative splicing alterations of VEGFA have also been already identified in lung cancer (40), enforcing the potential functional role of the aberrant transcript expression of this molecule. These few examples illustrate the impact that deeper functional studies might have to establish novel pathogenic mechanisms for SCLC at the molecular level.

Identification of such consistent alterations in splicing patterns leads us to wonder whether this could provide a new insight into targeted therapeutic strategies. For instance, the development of therapies with anti-sense oligonucleotides may be directed to target specific aberrant spliced transcripts in lung cancers as an alternative to small molecule inhibitors used until now (41). Therapies with anti-sense oligonucleotides have been successfully applied to other diseases and now they are starting to be seen as new therapeutic options in some cases of cancers that have well defined splicing alterations (42).

However, the association between the identified splicing alterations and specific SCLC oncogenic mechanisms is hindered by the lack of data from normal samples of SCLC patients. Since they exist many inherent tissue differences between SCLC and other lung tumours, in this work we cannot discriminate the real source of the altered isoform

expression. For this reason, it will be key to perform similar analyses with paired normal and tumour samples of SCLC as it has been done before for other cancers types (30).

In conclusion, we have been able to identify several genes that show switches in transcript isoforms in SCLC samples. These candidate genes need to be further evaluated in order to explore their functional implications and their potential role as prognostic factors or therapeutic targets.

Acknowledgements

We would like to thank Dr Jun Yokota and Dr Takashi Kohno for providing us data for SCLC and LADC samples. I also want to specially thank my supervisor, Dr Eduardo Eyras, for being patient and helpful when solving my doubts. Finally, I want to mention all the colleagues that have contributed to this work and did not hesitate in helping me or providing me useful advice.

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer Statistics , 2014. CA Cancer J Clin. 2014;64(1):9–29.
2. Cancer Research UK [Internet]. [cited 2015 Jun 5]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#heading-One
3. Blanpain C. Tracing the cellular origin of cancer. Nat Cell Biol. 2013;15(2):126–34.
4. Leslie KO, Colby T V. Pathology of lung cancer. Curr Opin Pulm Med. 1997;3(4):252–6.
5. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong K-K. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014;14(8):535–46.
6. Roy SH, Heymach J V, Lippman SM. Lung cancer. N Engl J Med. 2008;359(1-2):1367–80.
7. TCGA: The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543–50.
8. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res. 2012;72(1):100–11.
9. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. Lancet Oncol. 2011;12(2):175–80.
10. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med. 2010;362(25):2380–8.
11. Kwak EL, Bang Y-J, Camidge DR, Shaw AT, Dezube BJ, Jänne PA, et al. Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. N Engl J Med. 2010;363:1693–703.
12. Yokota J, Shiraishi K, Kohno T. Genetic basis for susceptibility to lung cancer: Recent progress and future directions. Adv Cancer Res. United States; 2010;109:51–72.
13. Iwakawa R, Kohno T, Kato M, Shiraishi K, Tsuta K, Noguchi M, et al. MYC amplification as a prognostic marker of early-stage lung adenocarcinoma identified by whole genome copy number analysis. Clin Cancer Res. 2011;17(6):1481–9.
14. Li F, Feng Y, Fang R, Fang Z, Xia J, Han X, et al. Identification of RET gene fusion by exon array analyses in “pan-negative” lung cancer from never smokers. Cell Res. 2012;22(5):928–31.
15. Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T, et al. KIF5B-RET fusions in lung adenocarcinoma. Nat Med. 2012;18:375–7.

16. Chen J, Weiss WA. Alternative splicing in cancer: implications for biology and therapy. Oncogene. 2014;34:1–14.
17. Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene. 2014;33:5311–8.
18. Amin EM, Oltean S, Hua J, Gammons MVR, Hamdollah-Zadeh M, Welsh GI, et al. WT1 Mutants Reveal SRPK1 to Be a Downstream Angiogenesis Target by Altering VEGF Splicing. Cancer Cell. 2011;20:768–80.
19. Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell. 2013;52:720–33.
20. Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. Mol Cell Biol. 2013;33(2):396–405.
21. Izquierdo JM, Majós N, Bonnal S, Martínez C, Castelo R, Guigó R, et al. Regulation of fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol Cell. 2005;19(4):475–84.
22. Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, et al. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature. 2011;480:387–90.
23. Das S, Anczuków O, Akerman M, Krainer AR. Oncogenic Splicing Factor SRSF1 Is a Critical Transcriptional Target of MYC. Cell Rep. 2012;1(2):110–7.
24. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol. 2007;14(3):185–93.
25. Ahlander J, Bosco G. The RB/E2F pathway and regulation of RNA processing. Biochem Biophys Res Commun. 2009 Jul 3;384(3):280–3.
26. Kikuchi S, Iwai M, Sakurai-Yageta M, Tsuboi Y, Ito T, Maruyama T, et al. Expression of a splicing variant of the CADM1 specific to small cell lung cancer. Cancer Sci. 2012;103(6):1051–7.
27. Coulson JM, Edgson JL, Woll PJ, Quinn JP. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: A potential role in derepression of neuroendocrine genes and a useful clinical marker. Cancer Res. 2000;60:1840–4.
28. Shimojo M, Shudo Y, Ikeda M, Kobashi T, Ito S. The Small Cell Lung Cancer-Specific Isoform of RE1-Silencing Transcription Factor (REST) Is Regulated By Neural-Specific Ser/Arg Repeat-Related Protein of 100 kDa (nSR100). Mol Cancer Res. 2013;11(10):1258–68.

29. Bourdon J, Fernandes K, Murray-zmijewski F, Liu G, Diot A, Xirodimas DP, et al. p53 isoforms can regulate p53 transcriptional activity. Genes Dev. 2005;19:2122–37.
30. Sebestyén E, Zawisza M, Eyras E. Recurrent alternative splicing isoform switches in tumor samples provide novel signatures of cancer. Nucleic Acids Res. 2015 Jan 10;43(3):1345–56.
31. Iwakawa R, Kohno T, Totoki Y, Shibata T, Tsuchihara K, Mimaki S, et al. Expression and clinical significance of genes frequently mutated in small cell lung cancers defined by whole exome/RNA sequencing. Carcinogenesis. 2015;36:616–21.
32. TCGA: The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
33. Li B, Ruotti V, Stewart RM, Thomson J a., Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2009;26(4):493–500.
34. Geman D, D'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004;3.
35. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005;21(20):3896–904.
36. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10.
37. Shimizu Y, Mullins N, Blanchard Z, ElShamy WM. BRCA1/p220 loss triggers BRCA1-IRIS overexpression via mRNA stabilization in breast cancer cells. Oncotarget. 2012;3(3):299–313.
38. Potiron VA, Roche J, Drabkin HA. Semaphorins and their receptors in lung cancer. Cancer Lett. 2009;273:1–14.
39. Mancino M, Ametller E, Gascón P, Almendro V. The neuronal influence on tumor progression. Biochim Biophys Acta. 2011;1816:105–18.
40. Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, et al. Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol. 2011;31(1):138–50.
41. Wood SL, Pernemalm M, Crosbie P, Whetton AD. Molecular histology of lung cancer : From targets to treatments. Cancer Treat Rev. Elsevier Ltd; 2015;41:361–75.
42. Bennett CF, Swayze EE. RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. Annu Rev Pharmacol Toxicol. 2010;50:259–93.

Supplementary material

Samples	Genes	Num.
Cell lines	ACADVL, AGPAT3, APOE, ARRB2, ASPH, ATPAF1, ATRIP, C3orf58, CAB39, CDK19, CREM, CRLS1, CRNDE, CSNK1G3, CUTA, DICER1, DYRK1B, EBAG9, FUS, H6PD, HDGF, HPCAL1, ITPRIPL2, MARCH8, MEST, METTL9, MIS12, MOGS, NEK6, NFKBIZ, NSUN5, PAPOLA, PDCD2, POLR1D, PPHLN1, PPP2R5A, PPP2R5E, PTGES2, RAD23B, RAP2C, RRAS2, RTN4, RUFY1, SYNCRIP, UBE2H, UHMK1, VEGFA, ZFAND5, ZNF331	49
Fresh tumour	APLP2, ARFIP1, ASPH, CAMKK2, CCDC117, CD8A, COPS7A, DAG1, DGKH, DICER1, ECM2, ENTPD1, FLT1, HIPK1, IFT20, KMT2E, MARCH8, MFAP3, MMP19, NCOA7, NFIB, NUMA1, PPCS, PPHLN1, PPP1R12B, PPP2R4, PPP2R5E, PTPRC, PUM2, RBPJ, RBPMS, RCAN1, RERE, RGS3, RGS5, RRAS2, SHROOM4, SLC4A7, TYMP, UBE2V1, UHMK1, USP10, USP38, USP4, ZBTB20, ZFHX3	46

Supplementary table S1. Genes that appear in at least 7 out of the 8 rankings from analyses of either SCLC fresh tumour or cell line samples against samples from LADC, LUAD tumour, mixed adenocarcinoma, LUSC tumour, mixed tumour, LUAD normal, LUSC normal and mixed normal as described in (Figure 2).

Samples	Genes	Num.
Tumour	AHCYL1, ASPH, C6orf48, CCDC117, CDK19, CSNK1G3, DICER1, ITGAV, MARCH8, MIS12, PPHLN1, PPP2R5E, RRAS2, UHMK1	14
Normal	AGER, AK3, APLP2, APOLD1, ARL6IP4, ASPH, C2CD5, CRLS1, DAG1, DICER1, DIXDC1, FGFR2, FLNB, GNAI2, JAM2, KANSL3, KDM2A, MACROD2, MARCH8, MEF2A, NUMA1, PPHLN1, PPP2R4, PRKAR1B, PTP4A2, RAD23B, RBPJ, RCAN1, RGS3, RPL32, RPS6KA1, RRAS2, SLC16A4, SORBS1, SULT1A2, UBE2V1, UHMK1, ZNF331	37

Supplementary table S2. Isoform changes shared by samples from SCLC fresh tumours and cell lines analysed against tumour (LADC, LUAD, LUSC, mixed adenocarcinoma and mixed tumour data sets) or normal samples (LUAD, LUSC and mixed normal data sets), as shown in (Figure 4).

SCLC cell lines against normal samples		
Description	**P-value**	**Num.**
Lung epithelial cell differentiation	2.82e-5	4
Lung cell differentiation	3.81e-5	4
Regulation of macromolecule metabolic process	1.4e-4	56
Regulation of cellular protein metabolic process	2.09e-4	30
Regulation of primary metabolic process	2.51e-4	55
Regulation of metabolic process	2.68e-4	63
Regulation of protein metabolic process	3.53e-4	31
Neuron projection morphogenesis	3.6e-4	7
Regulation of cellular metabolic process	4.21e-4	56
Clara cell differentiation	5.35e-4	2
Immune system process	5.66e-4	25
Cytoskeleton organization	6.05e-4	14
Regulation of signaling	7.48e-4	34
Regulation of cell communication	7.96e-4	34
Regulation of signal transduction	8.59e-4	31
Morphogenesis of a branching epithelium	9.08e-4	6
Positive regulation of biological process	9.25e-4	53
Protein modification by small protein removal	9.75e-4	5

Supplementary table S3. Results from functional enrichment analysis performed with resulting genes from comparing samples from SCLC cell lines against normal samples as described in (Table 3).

SCLC cell lines against tumour samples		
Description	**P-value**	**Num.**
Regulation of cellular process	1.01e-4	52
Lung epithelial cell differentiation	1.47e-4	3
Clara cell differentiation	1.73e-4	2
Lung cell differentiation	1.83e-4	3
Regulation of biological process	2.07e-4	53
Biological regulation	3.92e-4	54
Angiogenesis	4.96e-4	6
Cell differentiation	5.88e-4	16
Regulation of cellular metabolic process	5.89e-4	35
Regulation of macromolecule metabolic process	6.45e-4	34
Cellular protein modification process	8.03e-4	21
Protein modification process	8.03e-4	21
Commissural neuron axon guidance	8.51e-4	2
Protein modification by small protein removal	9.17e-4	4

Supplementary table S4. Results from functional enrichment analysis performed with resulting genes from comparing samples from SCLC cell lines against tumour samples as described in (Table 3).

SCLC fresh tumours against normal samples Description	P-value	Num.
Negative regulation of molecular function	4.68e-6	21
Negative regulation of catalytic activity	8.99e-6	18
Negative regulation of metabolic process	6.85e-5	33
Positive regulation of transcription from RNA polymerase II promoter in response to hypoxia	4.9e-4	2
Negative regulation of cellular metabolic process	5.9e-4	28
Regulation of phosphoprotein phosphatase activity	6.29e-4	4
Positive regulation of macromolecule metabolic process	6.87e-4	32
Regulation of RNA stability	8.03e-4	4
Regulation of mRNA stability	8.03e-4	4
Positive regulation of axonogenesis	9.01e-4	4
Forebrain ventricular zone progenitor cell division	9.72e-4	2

Supplementary table S5. Results from functional enrichment analysis performed with resulting genes from comparing samples from SCLC fresh tumours against normal samples as described in (Table 3).

SCLC fresh tumours against tumour samples Description	P-value	Num.
Regulation of axon extension	2.68e-4	4
RNA stabilization	3.15e-4	3
mRNA stabilization	3.15e-4	3
Regulation of extent of cell growth	5.26e-4	4
Positive regulation of axon extension	6.26e-4	3
Regulation of beta-amyloid clearance	7.17e-4	2
Artery morphogenesis	8.35e-4	3

Supplementary table S6. Results from functional enrichment analysis performed with resulting genes from comparing samples from SCLC fresh tumours against tumour samples as described in (Table 3).