


RESEARCH ARTICLE

Open Access



# Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response

Marta Lucchetta<sup>1</sup>, Isabelle da Piedade<sup>1</sup>, Mohamed Mounir<sup>1</sup>, Marina Vabistsevits<sup>1</sup>, Thilde Terkelsen<sup>1</sup> and Elena Papaleo<sup>1,2\*</sup> 

## Abstract

**Background:** Genomic initiatives such as The Cancer Genome Atlas (TCGA) contain data from -omics profiling of thousands of tumor samples, which may be used to decipher cancer signaling, and related alterations. Managing and analyzing data from large-scale projects, such as TCGA, is a demanding task. It is difficult to dissect the high complexity hidden in genomic data and to account for inter-tumor heterogeneity adequately.

**Methods:** In this study, we used a robust statistical framework along with the integration of diverse bioinformatic tools to analyze next-generation sequencing data from more than 1000 patients from two different lung cancer subtypes, i.e., the lung adenocarcinoma (LUAD) and the squamous cell carcinoma (LUSC).

**Results:** We used the gene expression data to identify co-expression modules and differentially expressed genes to discriminate between LUAD and LUSC. We identified a group of genes which could act as specific oncogenes or tumor suppressor genes in one of the two lung cancer types, along with two dual role genes. Our results have been validated against other transcriptomics data of lung cancer patients.

**Conclusions:** Our integrative approach allowed us to identify two key features: a substantial up-regulation of genes involved in O-glycosylation of mucins in LUAD, and a compromised immune response in LUSC. The immune-profile associated with LUSC might be linked to the activation of three oncogenic pathways, which promote the evasion of the antitumor immune response. Collectively, our results provide new future directions for the design of target therapies in lung cancer.

**Keywords:** Lung adenocarcinoma, Lung squamous cell carcinoma, Differential expression analysis, RNA-Seq, Co-expression, Soft clustering, Survival analysis, TCGA

## Background

Lung cancer is one of the most aggressive cancers, with a five-year overall survival of 10–15% [1]. Lung cancer can be classified into small cell lung cancer (SCLC) and non-SCLC (NSCLC), which account for 15 and 85% of all lung cancers, respectively. The main subtypes of NSCLC are divided into adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). Lung cancer is a

highly heterogeneous cancer type with multiple histologic subtypes and molecular phenotypes [2, 3].

Since 2015, the classification of lung tumors has been defined by cytology and histology [1, 4]. Despite the staining strategy to separate lung tumors into different classes, cases that are ambiguous at the immunohistochemical level are often reported and difficult to resolve. A proper differentiation between LUAD and LUSC determines eligibility for certain types of therapeutic strategies [5]. For example, some drugs are contraindicated for one of the two lung cancer types, such as Bevacizumab (Avastin) in LUSC [6]. It thus becomes crucial to discriminate among the two lung cancer types in a precise way.

\* Correspondence: [elenap@cancer.dk](mailto:elenap@cancer.dk)

<sup>1</sup>Computational Biology Laboratory, Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark

<sup>2</sup>Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark



Microarray technologies have been used to identify differentially expressed genes in lung cancer samples helping to pinpoint critical markers [7–10]. For example, Naval et al. identified a prognostic gene-expression signature of 11 genes, which was subsequently validated in several independent NSCLC gene expression datasets [9]. This pioneering study established the prognostic impact of changes in gene-expression for NSCLC patients. However, the markers identified in the study do not differentiate between LUAD and LUSC.

Gene or microRNA markers may be used, in principle, to distinguish between these two types of cancer [11–14]. In this context, single markers are unlikely to be sufficiently robust to discriminate between cancer subtypes due to the intrinsic heterogeneity of tumors. New methods have been developed for robust analyses of co-expression signatures in gene expression data [15–17]. Co-expressed modules are groups of genes that not only show high correlations of expression; they also encompass important information on the heterogeneity of phenotypes, cancer progression, or response to treatment [18–24]. We speculated that the integration of differential expression and co-expression analyses could be beneficial for the comparison of cancer (sub)types.

The Cancer Genome Atlas (TCGA) is a large genomic initiative in which more than 10 000 patients were profiled using six different platforms to identify cancer-related signatures [25–27]. TCGA provides a unique resource which can be re-analyzed for the discovery of cancer-related alterations or new biomarkers specific to certain cancer (sub)types. Among the next-generation sequencing (NGS) platforms available, RNA-Seq is a reliable approach for quantification of changes at the transcriptional level [28]. Lung cancer datasets for LUAD [29] and LUSC [30] are available in TCGA and account for more than 1000 samples overall. In parallel, the Recount2 initiative [31], which integrates GTEX (Genotype-Tissue Expression Project) [32] and TCGA, has recently allowed for an increase of healthy tissue samples for the comparison with tumor samples. Thus, the LUAD and LUSC TCGA datasets offer a suitable starting point for the identification of gene expression signatures that could discriminate between the two lung cancer types in terms of classification, diagnosis, and prognosis, along with to shed light on the underlying molecular mechanisms. These two TCGA datasets have been used either to identify general cancer signatures [10, 33–39] or to pinpoint signatures specific to only one of the lung cancer types [40–42].

Cline and colleagues [39] recently showed that there is a subset of 19 samples in the TCGA LUSC cohort that feature a LUAD-like gene expression profile. They labeled these samples ‘discordant LUSC’. Discordant LUSC samples are borderline for subtype classification,

and the similarity with LUAD is also modest. These findings were supported by the analyses on an alternative pre-processing of the TCGA datasets [38]. As such, it is essential to account for this information in the re-analysis of the TCGA lung cancer data to avoid misleading conclusions.

We aimed to closely compare LUAD and LUSC TCGA datasets using a robust statistical and bioinformatic framework (Fig. 1). In particular, we: i) identified a group of genes that are differentially expressed between LUAD and LUSC when compared to the normal samples; ii) assessed changes in the gene expression signature over cancer stages; iii) identified modules of differently co-expressed genes in the two lung cancer types and alterations in their transcriptional regulation; iv) predicted potential oncogenes, tumor suppressors or dual role genes for each type and, iv) evaluated if any of the LUAD- or LUSC-specific candidate genes had a prognostic potential. Overall, our study resulted in a subset of genes and pathways that could be used to discriminate among the two cancer types. Moreover, we identified candidate genes which are suitable for further functional/structural studies since they are poorly understood and potentially important as lung cancer markers or targets. Our data also provide a useful guide for future cellular studies using cancer cell lines, which reflects the LUAD or LUSC types.

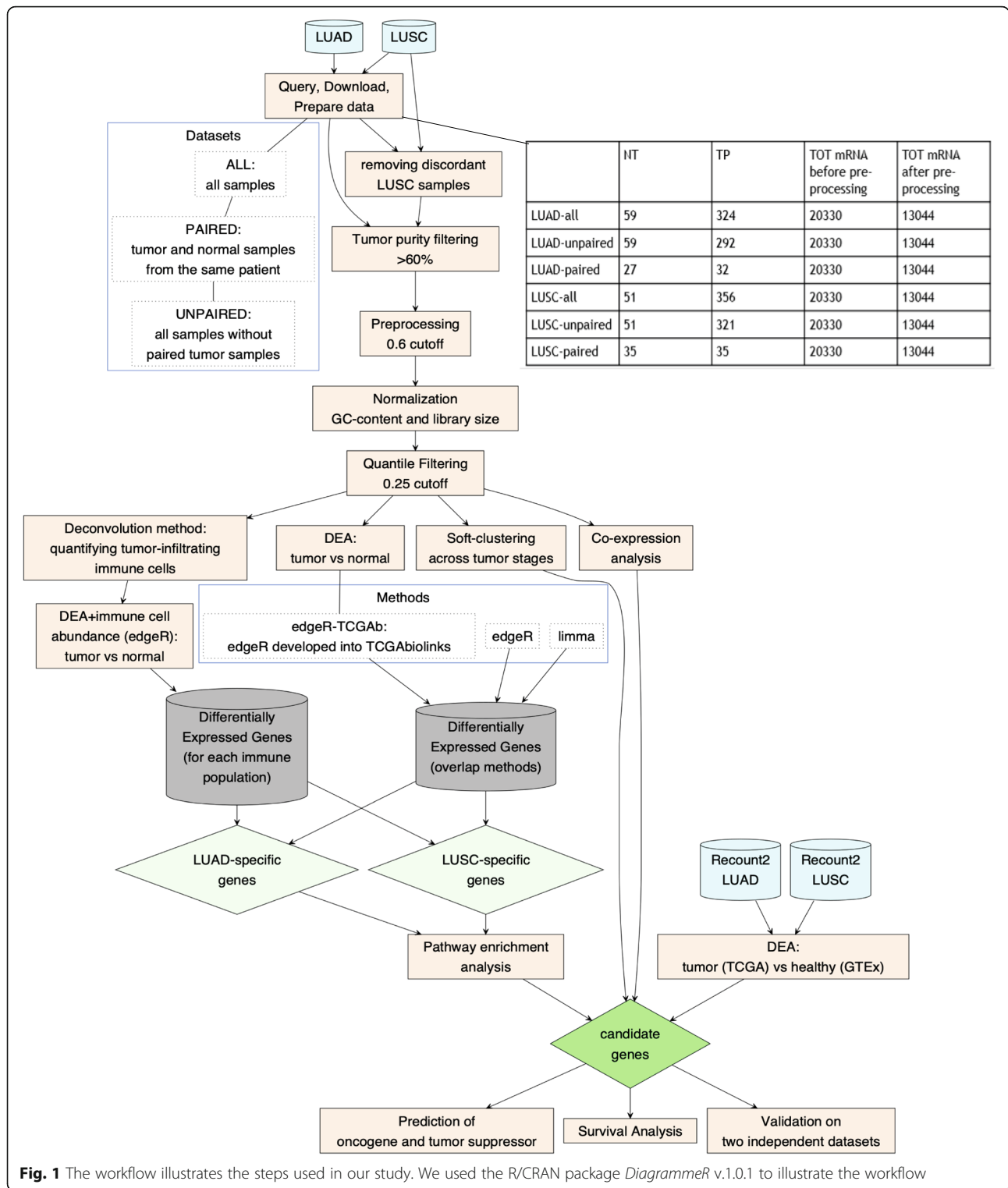
## Methods

### Pre-processing of RNA-Seq data from the Cancer genome Atlas (TCGA)

We downloaded and pre-processed level 3 legacy RNA-Seq data (RSEM count) for LUAD and LUSC with the *GDCquery* of the *TCGAbiolinks* Bioconductor/R package [43, 44]. The RNA-Seq data have been produced using the Illumina HiSeq 2000 mRNA sequencing platform.

We downloaded the data in October 2016 from the Genomic Data Common (GDC) Portal (<https://portal.gdc.cancer.gov>). An overview of the analyzed samples is reported in Fig. 1. We removed the 19 ‘discordant LUSC’ samples before analysis [38], along with samples with low tumor purity (< 60%) using a consensus measurement of tumor purity [45].

We then employed the *TCGAbiolinks* [43] function *GDCprepare* to obtain a *Summarized Experiment* object [46]. We removed outlier samples with the *TCGAanalyze\_Preprocessing* function of *TCGAbiolinks* using a Spearman correlation cutoff of 0.6. We normalized the datasets for GC-content [47] and library size using the *TCGAanalyze\_Normalization*. Lastly, we filtered the normalized RNA-Seq data for low counts across samples using the function *TCGAanalyze\_Filtering*. This step removed all transcripts with mean across all the samples less than



0.25 quantile of the mean. The pre-processed and processed datasets are available through our *GitHub* repository, along with the script to generate them ([https://github.com/ELELAB/LUAD\\_LUSC\\_TCGA\\_comparison](https://github.com/ELELAB/LUAD_LUSC_TCGA_comparison)).

**Differential expression analyses of TCGA datasets**

Differential expression analyses have been carried out using *edgeR* [48] and *limma* [49]. The analyses were performed using three different pipelines. One pipeline was based on *limma-voom* and the other two were *edgeR*-based. One of

the two *edgeR*-based pipelines was implemented in the *TCGAanalyze\_DEA* function which was incorporated into the first release of the *TCGAbiolinks* package (called *edgeR-TCGAb* in this study for sake of clarity).

We applied the *voom* transformation to RNA-Seq data analysis with *limma* [50] to estimate precision weights for each count, before performing the differential expression analysis. In *edgeR* pipelines, we used the *GLM* (Generalized Linear Models) approach. Both the *limma* and *edgeR* pipelines allowed to include an experimental design with multiple factors.

In the *limma* and *edgeR* pipelines, the design matrix includes: conditions (tumor vs normal), the patient information when a paired dataset was used, and batches for the other datasets (*unpaired* and *all*). We corrected for the TSS (Tissue Source Site; the center where the samples are collected) as source of batch effect in the *edgeR* and *limma-voom* DEA pipelines. In contrast, *edgeR-TCGAb* implemented a simple function for DEA which did not account either for batch corrections or patient information. In all the DEA analyses, we defined as a cutoff a log fold change (logFC)  $\geq 1$  or  $\leq -1$  to retain significant DE genes, along with a False Discovery Rate (FDR) cutoff of 0.01.

During the analyses, we tested two variations of the *limma-voom* DEA pipeline: i) using the same design matrix for *voom* and *lmFit* functions and ii) using the entire *voom* object in the steps following the *voom* transformation and not only the log2-transformed data. These adjustments provide a more correct approach to DEA but did not make any difference on our final conclusions. The corresponding scripts are also reported in our *GitHub* repository.

The overlap between the DE genes identified by each pipeline and for each different curation of the dataset was evaluated with the *UpSetR* package [50].

#### Curation and differential expression analyses of unified GTEx and TCGA LUAD and LUSC datasets

We used the unified dataset integrating the GTEx [32] cohort of healthy samples and TCGA data as provided by the *Recount2* protocol [31]. We employed the *TCGAquery\_Recount2* function of *TCGAbiolinks* v2.8 to query the GTEx and TCGA unified dataset for lung cancer [51]. We filtered the data for tumor purity with a threshold of 60% as we did for the TCGA dataset and removed the LUSC discordant samples.

Since *recount2* barcodes were updated to the Universally Unique Identifier (UUID), we convert them to filtered TCGA barcodes with the *TCGAAutils* package, so that we could apply the pre-processing steps with *TCGAbiolinks*. The mapping between the TCGA barcodes and the new UUIDs was obtained by extracting the GDC case identifiers. We analyzed 374, 355 and, 393 samples for GTEx, LUAD

and, LUSC, respectively in the unified datasets. After the filtering steps and preparation of the unified datasets for LUAD and LUSC, only protein-coding genes were retained using the *biomaRt* Bioconductor/R package [52–54]. We carried out GC-content normalization and quantile filtering as described above. We converted the ENSEMBL identifiers into gene names through the information in the *SummarizedExperiment* object. The DEA was carried out with the *limma-voom* method according to the pipeline described in the previous section.

#### Analysis of the tumor microenvironment

To appreciate the differences between LUAD and LUSC immune landscapes, we estimated the abundance of populations of tissue-infiltrating immune cells using the R package *MCP-counter* [67]. We used the *MCPcounter.estimate* function to estimate the abundance of immune cell populations (T cells, cytotoxic lymphocytes, B cell and, monocytic lineage, myeloid dendritic cells, neutrophils) and non-immune stromal populations (endothelial cells and fibroblasts) for each sample. In particular, the gene matrix obtained after the pre-processing steps was used as input for the analysis.

For each cell population, we divided the samples into four groups (see *GitHub* repository): i) very-low (population from 0 to the minimum value of the distribution); ii) low (from the minimum to the first quartile); iii) medium (from the first to the third quartile); and iv) high (from the third quartile to the maximum value) abundance. We incorporated this information into the design matrix for DEA with *edgeR* to identify the DE genes for each cell population and each lung cancer type in comparison with the normal samples. These different sets of DEAs were compared to the consensus DEA for the selection of the candidate genes and to assess the robustness of the pathway-enrichment and GO-enrichment analyses (see [Pathway enrichment analyses](#) Section).

#### Soft-clustering analysis

We performed gene clustering for the LUAD and LUSC datasets *paired* and *all* according to the normal tumor (NT) and four clinical stages of cancer, i.e., stages I, II, III, and IV using the *Mfuzz* package version 2.36.0 [55]. *Mfuzz* uses a fuzzy c-means algorithm based on the iterative optimization of an objective function to minimize the variation of objects within the clusters. The fuzzy c-means algorithm is robust to noise and avoids a priori pre-filtering of genes [56]. We assessed the longitudinal evolution of the mean of expression in the LUAD and LUSC clusters in the normal samples and along the four stages of tumor progression.

The dataset with all the samples were used for the soft-clustering analyses and a consensus matrix containing all the gene expression for LUAD and LUSC was built. We collected all barcodes corresponding to NT samples using the *TCGAbiolinks* function *TCGAquery\_SampleTypes* for each lung subtype to identify NT samples in the gene matrices. We mapped the tumor samples to their stages from the LUAD and LUSC clinical datasets using their barcodes. We filtered out the samples with “not reported” stage status. We computed the mean gene expression value per tumor stage and NT for LUAD and LUSC. We defined a maximum number of clusters equal to six for the fuzzy *c*-means clustering. The centroid clustering step results from a weighted sum of all cluster members and shows the overall gene expression pattern in each cluster. The membership values indicate how well a gene is represented by its cluster. Low values illustrate a poor representation of the gene by the cluster centroid. Large values point to a high correlation of the expression of the gene with the cluster centroid. In each cluster, genes are represented with color lines corresponding to their cluster membership  $m > 0.56$ . We selected this cutoff empirically to be stricter than the default value of 0.5. The membership values are color-encoded in the plots generated by *mfuzz.plot*.

#### Pathway enrichment analyses

We used the *ReactomePA* version 1.18.1 R/Bioconductor package to perform the Reactome-based Pathway Analysis [57]. We employed the *enrichPathway* function of *ReactomePA* to retrieve the enriched pathways in the DE gene set or in the list of genes from the soft clustering. We used the entire gene matrix after pre-processing as a background. An adjusted *p*-value cutoff of 0.05 was set, and the analysis was done by separating the up- and down-regulated genes for each dataset (*all* and *paired*) and lung cancer subtype. In addition, all the gene symbols were converted to their corresponding ENTREZ IDs provided by the *SummarizedExperiment* object (*GDCprepare* function output). We used *clusterProfile* [58] to illustrate the results of the pathway enrichment analyses.

#### Gene ontology enrichment analysis

To identify biological functions in LUAD and LUSC DE gene sets, we carried out a Gene Ontology (GO) classification, which included the following categories: biological process, cellular component and molecular functions [59].

We performed GO functional enrichment analysis for the DE gene set using the *topGo* R/Bioconductor package. We provided both DE and background genes lists separating up- and down-regulated genes. We used the same background used in the [Pathway enrichment](#)

[analyses](#). The GO results for the biological processes were represented in circular plots with the *GOplot* R package [60].

#### Co-expression network analyses

We used the LUAD and LUSC datasets upon filtering and after *voom* transformation (see [Differential expression analyses of TCGA datasets](#) section) to carry out modular co-expression analyses with the *CEMiTool* Bioconductor/R package [15] using the protocol suggested by the developers. We performed pathway enrichment analyses and protein-protein network analyses with the pre-built functions of *CEMiTools*. As a reference for protein-protein interactions, we used the *Interologous Interaction Database I2D* version 2.9. [61].

#### Survival analysis

We performed a survival analysis using the *survival* R package version 2.41–3. We used Cox regression [62] to estimate differences in survival between patients with low and high expression levels of our candidate genes. For each cancer type, tumor samples were extracted and separated by gene expression levels according to lower and upper percentile (25th and 75th, respectively). If the gene expression level of a specific gene in a certain sample was lower than the 25th percentile, the corresponding sample was labeled as *low*. Samples with the gene expression level greater than the 75th percentile were labeled as *high*. In cases of tumor duplicates (i.e., tumor samples from the same patient), we used the mean for the analysis. The clinical data were downloaded using the *GDCquery\_clinical* function of *TCGAbiolinks*. We used only barcodes for which information regarding the last follow-up or death time of the patient was available.

Cox regression analyses were performed using the *coxph* function. Cox regression allows to account for additional explanatory variables, such as age at diagnosis, gender, and tumor stage. Before performing Cox regression, we tested the proportional assumption using the *cox.zph* function, and we retained only the genes which satisfy this test (11 and 13 genes for LUAD and LUSC, respectively). The *p*-values of each variable were corrected using the Benjamini and Hochberg (BH) method [63].

#### In silico validation of the candidate genes on independent cohorts

To validate our candidate genes, we selected two microarray studies that include both LUAD and LUSC samples. The first study contains 139 and 21 samples for LUAD and LUSC, respectively [64]. The second dataset (GEO accession number: GSE33532) is composed of ten and four samples for LUAD and LUSC, respectively [65]. At first, the probe sets were converted to gene names

using the *gconvert* function of the *gProfileR* package version 0.6.6 [66] and the non-converted probes were removed. Since multiple probe sets can identify the same gene, we collapsed them to obtain unique matches with the *collapseRows* function implemented in the *WGCNA* package version 1.63 [16]. We performed hierarchical clustering with the complete method and euclidean distance and visualized the results as heatmap with the *heatmap.2* function in the *gplots* package version 3.0.1.

## Results

### Curation and description of the datasets used in the study

The datasets for our analyses were curated to remove the LUSC discordant samples, along with samples with a low tumor purity and outlier samples with correlation lower than 0.6. The number of samples and genes retained for the analyses are reported in Fig. 1.

We performed differential expression analyses (DEAs) to identify a subset of differentially expressed genes in the two TCGA lung cancer datasets LUAD and LUSC tumor primary (TP) with respect to the normal (NT) samples. A clear consensus on the best DEA approaches for RNA-Seq data does not exist yet. Different DEA methods could provide different results [49, 68–71]. We employed three pipelines for DEA of LUAD and LUSC and generated a consensus list of DE genes (see [Method](#) section).

In addition, we curated three different datasets (*paired*, *unpaired*, and *all*) for each cancer type and the results are discussed in the Additional file 1: Text S1. A comparison of the DE genes obtained from the analysis of each of the three datasets allowed us to evaluate the impact of the different curations in terms of sample size or sample pairs.

### The usage of a too simplistic design in the DEA protocol has marked effects on the DEA results

At first, we assessed the influence of using different definitions of the dataset for DEA analyses. We compared the results of DEA carried out with a certain method, i.e., *limma* or *edgeR* or *edgeR-TCGAb*, on the three different datasets (*paired*, *all*, and *unpaired*) for each of the two cancer types (LUAD and LUSC) (Figs. 1 and 2, and Additional file 1: Text S1). In our LUAD analyses, we used the *paired* dataset with 32 tumor and 27 normal samples, along with the *all* dataset with 324 tumor and 59 normal samples. For LUSC, the *paired* dataset contained 35 tumor and 35 normal samples, and the dataset *all* 356 tumor and 51 normal samples. In both cases, the *paired* datasets were small subsets of the corresponding full datasets. According to the comparison described in Additional file 1: Text S1, we focused on the dataset

with all the normal and tumor samples for the following analyses.

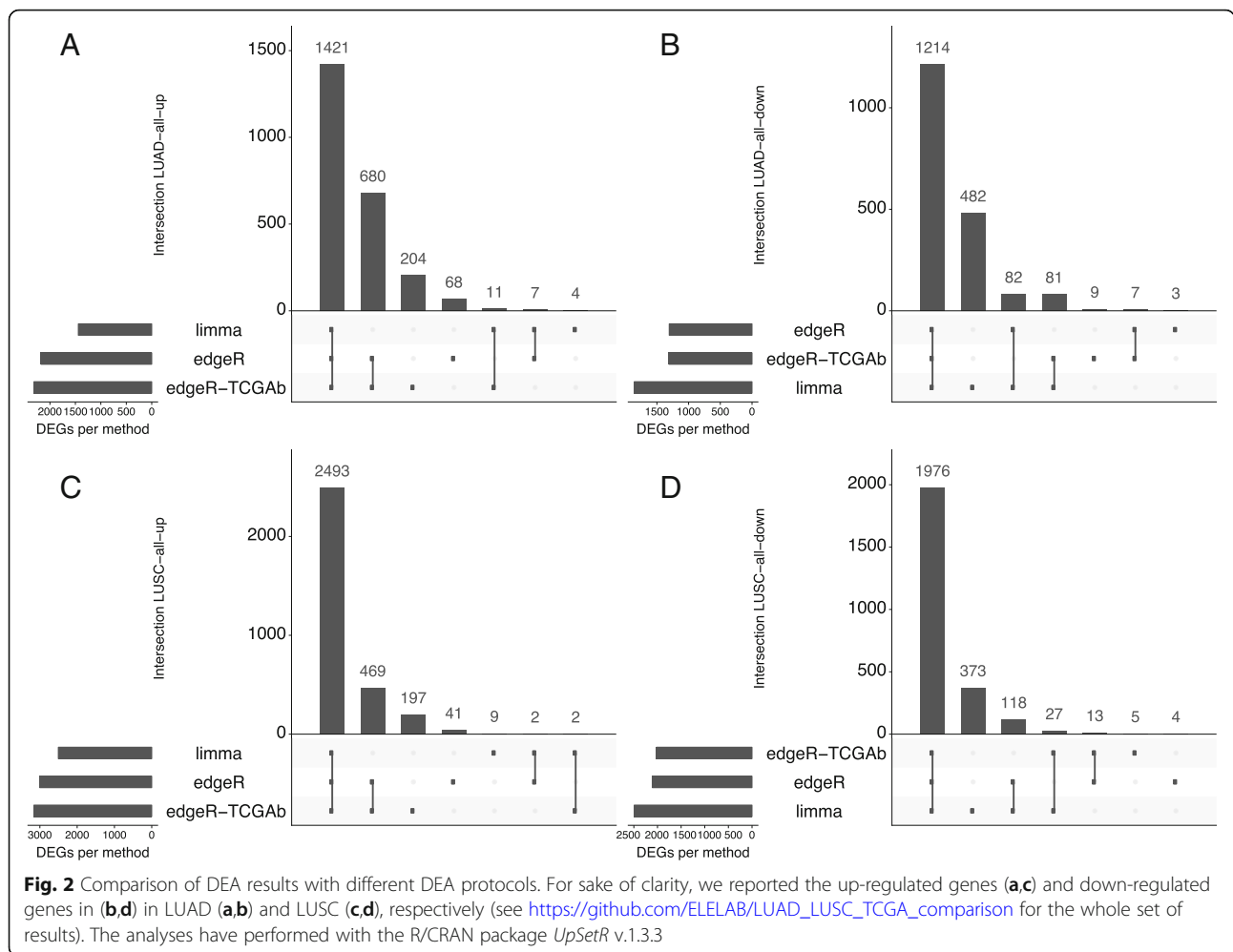
*Limma* resulted in the most stringent approach for up-regulated genes (Additional file 5: Table S1, Fig. 2a-c). Inversely, *limma* provided a large number of down-regulated gene (more than 300), which were not identified by the *edgeR* pipelines (Fig. 2b-d).

*EdgeR-TCGAb* featured a subset of up-regulated genes which were not identified by the other two methods (Fig. 2a-c). In the case of paired samples, this behavior can be explained by the fact that the *edgeR-TCGAb* pipeline does not correct for patient-specific effects, which are likely more important when the normal and tumor samples are matched. Moreover, *edgeR-TCGAb* DEA pipeline does not include batch corrections, which we included within the design matrix in the other two DEA pipelines. We had a closer look at the 820 and 619 up-regulated genes identified only by the *edgeR-TCGAb* pipeline in LUAD and LUSC (Additional file 2: Figure S1). Most of the discordant genes have either logFCs close to or below 1 or FDR values close to 0.01. Moreover, *edgeR-TCGAb* tends to overestimate the logFC values. We also noticed that there was a small number of cases in which *edgeR-TCGAb* assigned an opposite directionality, i.e., the genes were down-regulated according to the other two methods. Specifically, this set of genes included: DUOXA2, IGFALS and, KLK14 in LUAD, as well as EDN3, GF11B, MYH15, PEG3, and PENK in LUSC. We searched for each of the non-congruent genes in the *IGDB.NSCLC* database [72], which is a collection of genes that are altered in NSCLC. We did not consider hits for which the probe sets were reported with mapping problems or the fold change was lower than 2. We observed that only MYH15 was up-regulated in one of the LUSC studies at *IGDB.NSCLC*, while the other genes listed above were down-regulated, supporting the *edgeR* and *limma* results from our study.

The results of our analyses, thus, raised concerns about the accuracy of the original *TCGAbiolinks* DEA pipeline with *edgeR* (*edgeR-TCGAb*) especially when paired samples are analyzed, highlighting a need for a different DEA design within the R/Bioconductor package. This design should include functions for proper batch corrections and corrections for patient-specific effects, which we recently implemented in *TCGAbiolinks* [51]. Overall, 60–80% of the DE genes are in common among the three DEA methods, suggesting that their integration may allow for the removal of genes with borderline results to define a robust signature of LUAD- and LUSC-specific genes.

### Identification of LUAD- and LUSC-specific differentially expressed genes

We selected the datasets containing all the samples to maximize the sample size. As an additional control of



our analyses, we carried out DEA on the LUAD and LUSC *unified* datasets from the recent *Recount2* initiative [31]. In the *Recount2* platform, TCGA data were integrated with the normal GTEx [32] samples. This integration increases the pool of available normal samples for the comparison for a total of 374 healthy samples. Moreover, *Recount2* provides a genuine source of healthy tissue samples to compare with lung tumors, e.g., not only the normal adjacent tissues that are available in the TCGA. The list of DE genes for LUAD and LUSC for the *unified* datasets are reported in our *GitHub* repository.

We employed a consensus approach, in which we defined as DE genes in LUAD and LUSC only those found by all the three DEA approaches (i.e., the intersects in each of the overlap diagrams similar to the ones reported in Fig. 2). We then compared the up- and down-regulated genes in LUAD with the ones of LUSC. To identify gene signatures that can differentiate between the two lung cancer types, it is not sufficient that the genes are differentially expressed with

respect to the normal samples. One also needs to verify that they are not up (or down) regulated in both the cancer types.

We retained 337 and 1451 genes up-regulated, as well as 165 and 956 down-regulated genes in LUAD and LUSC, respectively (reported in our Github repository).

To verify that some of the differences observed for LUAD and LUSC DE genes did not come from differences in the composition of the tumor microenvironment between the two cancer types, we carried out additional 18 DEAs. In these DEA, we corrected for the populations of the different cellular infiltrations, which have been estimated using a deconvolution method (see [Analysis of the tumor microenvironment](#) section and Github repository for the lists of DE genes in different comparisons).

Interestingly, we identified a small subset of genes which were up-regulated in LUAD but down-regulated in LUSC (MUC5B, HABP2, MUC21, and KCN5) or vice-versa (CSTA, P2RY1, and ANXA8) in all or the majority of the DEA comparisons.

We performed pathway enrichment analysis for the up- and down-regulated unique genes of LUAD and LUSC from the consensus DEA comparison, along with for each of the DEA analyses correcting for infiltration from other cellular populations. We only retained the genes and pathways common to the different DEA calculations. The analyses revealed that pathways related to O-linked glycosylation of mucins was enriched for the up-regulated genes of LUAD and down-regulated genes of LUSC, respectively (Table 1). This suggests that the proteins involved in this pathway could play an important role in discriminating between the two lung cancer types. Of particular interest is a group of mucins (MUC5B, MUC15, MUC16, and MUC21), along with enzymes involved in their modifications. Additionally, we noticed that genes involved in the complement system (C3, C5, CD55, and CFI) were down-regulated in LUSC. GO-enrichment analysis on LUAD DE genes also confirmed the importance of up-regulation of processes related to O-linked glycosylation, along with cellular adhesion (Additional file 3: Figure S2).

#### Clustering of genes in LUAD and LUSC across tumor stages

We aimed to identify a subset of specific and interrelated genes which, as an ensemble, could be more effective than single markers in discriminating between LUAD and LUSC. For this purpose, DEA alone is not sufficient. We, therefore, analyzed the molecular signatures both using soft-clustering approaches over the stages of tumor progression and implementing weighted co-expression analyses.

We applied a soft-clustering approach [55, 56] to separate LUAD and LUSC genes into clusters based on their changes in gene expression in different tumor

stages [73], allowing us to identify six clusters with different signatures (Fig. 3a-c).

Clusters 2 and 5 of LUAD (Fig. 3a), along with clusters 1 and 5 of LUSC (Fig. 3c) revealed a general up-regulation of genes along all stages. The two up-regulated clusters of genes in LUAD showed enrichment in transcriptional regulation of p53, cellular response to stress and, mitosis. In LUSC, mitosis was also up-regulated, together with other processes here among translation, cell cycle, ER, Golgi and COPI transports (Fig. 3b and d). In contrast, clusters 1 and 6 of LUAD (Fig. 3a), and clusters 2 and 6 of LUSC (Fig. 3c) featured a general down-regulation when the four tumor stages were compared to the normal samples, with no clear enrichment in biological pathways.

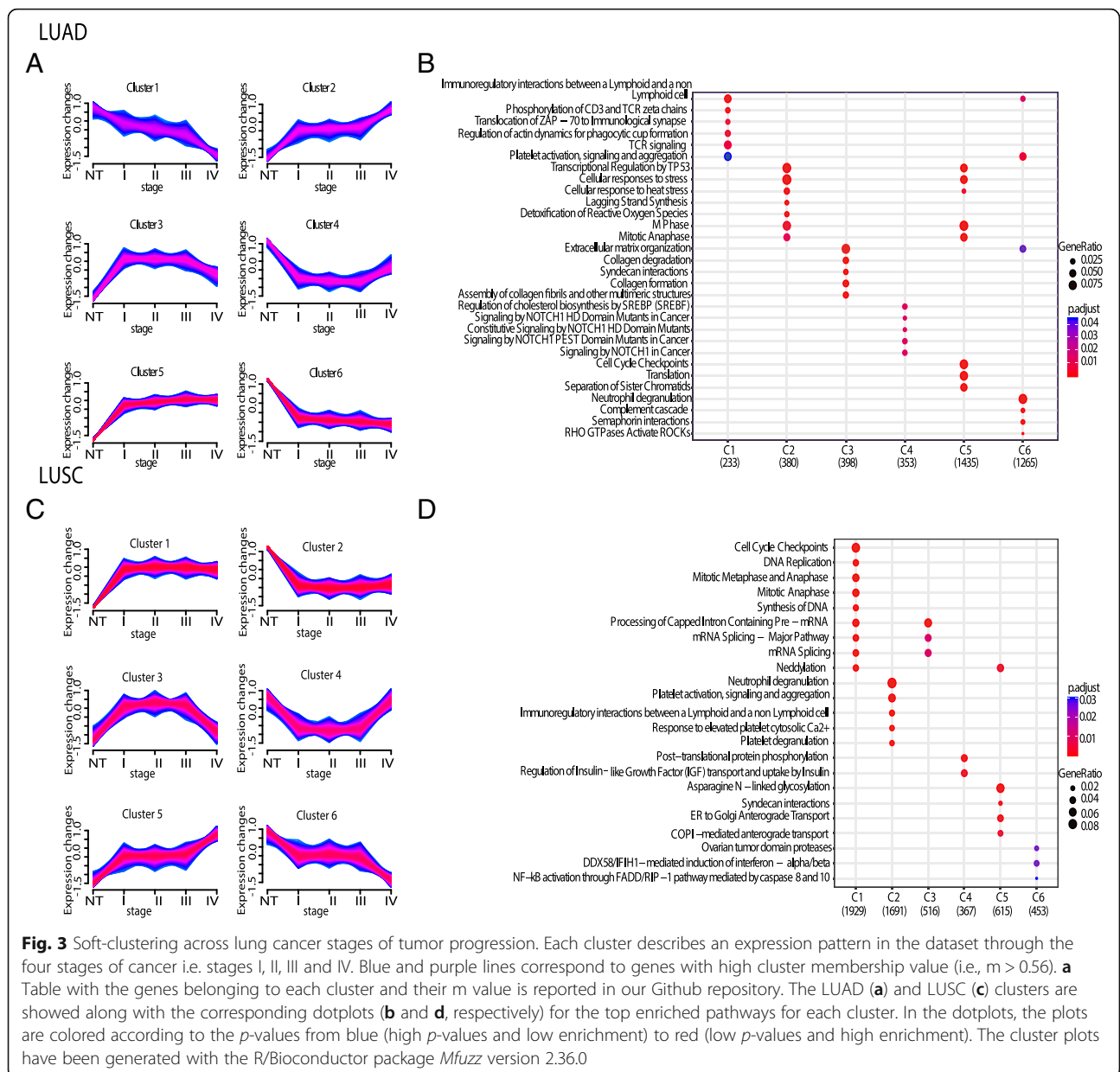
We extracted the genes that showed a trajectory of up-regulation across stages in one cancer type and down-regulation in the other, similarly to what we previously did for the DEA results. We identified a group of 46 genes which were up-regulated in LUAD but down-regulated in LUSC. 72 genes were down-regulated in LUAD and up-regulated in LUSC. The soft-clustering comparison provided an additional list of gene candidates of which MUC5B, CSTA, P2RY1 and, NTRK2 were shared between the soft-clustering and the DEA.

Clusters 3 of both LUAD and LUSC (Fig. 3a and c) featured a signature in which the genes were up-regulated at the early stages, but they decreased again at late stages (i.e. stage IV). Clusters 4 showed the opposite trend, i.e. a down-regulation at early stages but increase at late stages (Fig. 3a and c). These patterns may be indicative of dual-role genes [74]. The enriched processes were different for these genes in LUAD and LUSC. The dual-role of LUAD was associated with extracellular matrix organization, whereas in LUSC with mRNA splicing and mRNA processing (Fig. 3b and d). Expression

**Table 1** Pathway enrichment analysis with *ReactomePA*. Only the results relevant to the comparison of O-glycosylation, immune response and complement pathways are reported. For the full list of results, one could refer to our Github repository for the project. We reported only the pathways and genes identified by the initial consensus DEA and the DE genes from the DEA comparisons accounting for the tumor microenvironment. The range of FDR values identified for that pathway from the different enrichment analyses are reported as a reference

Pathway ID	Description	FDR	Gene IDs
913709	O-linked glycosylation of mucins	0.007–0.04	LUAD (up): B3GNT6/MUC16/MUC21/MUC5B
913709	O-linked glycosylation of mucins	0.01–0.07	LUSC (down): B3GNT7/B3GNT8/GALNT10/GALNT5/MUC1/MUC15/MUC21/MUC5B/ST6GALNAC4
977068	Termination of O-glycan biosynthesis	0.01–0.07	LUAD (up):MUC16/MUC21/MUC5B
5173105	O-linked glycosylation	0.01–0.05	LUAD (up): B3GNT6/ MUC16/MUC21/ MUC5B
5173105	O-linked glycosylation	0.01–0.03	LUSC (down): B3GNT7/B3GNT8/GALNT10/GALNT5/MUC1/MUC15/MUC21/MUC5B/ST6GALNAC4/THBS1
977606	Regulation of Complement cascade	0.0002–0.06	LUSC (down):C3/C5/CD55/CFI
166658	Complement cascade	0.0008–0.06	LUSC (down): C3/C5/CD55/CFI





of dual-role genes may, for example, be unwanted by cancer cells in early tumor stages, whereas the same genes become essential at later stages of tumorigenesis, providing the cancer cells with a functional advantage or resistance to chemotherapy [37].

**Prediction of oncogenes and tumor suppressor genes in LUAD and LUSC**

Genes that are up- or down-regulated and are also known to be oncogenes or tumor suppressors, respectively, are of great interest in cancer.

We, therefore, carried out a prediction of potential tumor suppressor genes (TSGs) and oncogenes (OGs) using *Moonlight* [37], which employs gene expression

signatures and biological pathways to identify potential TSGs and OGs. This analysis was used to integrate and expand the information available on TSGs and OGs through the curated data from *TSGene* (*TSGDB*) [75], *ONGene* [76], and *COSMIC* [77].

At first, we were interested in evaluating which of the up- and down-regulated genes that discriminate between LUAD and LUSC are known or predicted to be OGs (up-regulated genes) or TSGs (down-regulated genes).

We thus identified 24 potential TSGs and 146 OCGs for LUAD, while we obtained 22 TSGs and 456 OCGs for LUSC with *Moonlight*. The details and the full list of genes are reported in our *GitHub* repository. Only 31 predicted OCGs and no TSGs were common between

LUAD and LUSC. Intriguingly, IL6 and KRT23 were predicted as OCGs in LUAD but TSGs in LUSC, highlighting that these genes could deserve attention in future studies. IL6 is of interest because of its role in the immune response and the complement system [78] and its down-regulation in LUSC. We recently identified IL6 as the only down-regulated cytokine in breast cancer samples using cytokine assays [79]. Future studies on naïve tumor samples or LUSC and LUAD cellular models, where the IL6 gene can be modulated by over-expression or silencing, could shed light on its role within the two lung cancer types [80].

### Co-expression signatures in LUAD and LUSC

As stated above, we sought gene expression signatures to discriminate between LUAD and LUSC types, or interesting targets for each cancer type. For this reason, we also carried out a gene co-expression analysis to identify different modular gene co-expression networks in LUAD and LUSC.

In LUSC, we identified six modules (Fig. 4a). M1 was enriched in proteins for organization and assembly of the cell and gap junctions, including gap junction proteins, like the up-regulated hub proteins GJB5, keratin type II proteins and protein channels activated by chloride. M2 was enriched in proteins for glutathione conjugation and response to redox stress, such as the up-regulated hub proteins sulfiredoxin-1 protein and the oxidative stress-induced growth inhibitor OSGIN1. M3 included extracellular matrix organization and collagen-related proteins. M4 was enriched in interferon signaling, cytokine signaling in immune response, with a down-regulation of HLA genes. M5 was not associated with any annotated cellular pathway, whereas M6 was enriched in proteins that regulate the complement cascade.

We identified four modules in LUAD (Fig. 4b): M1 which was enriched in extracellular matrix organization proteins and regulation of complement cascade; M2, which was enriched in interferon and cytokine signaling and, M3 which included collagen-related genes and proteins for extracellular matrix organization. Notably, M3 was the only module that included hubs which were conserved among LUAD and LUSC. M4 of LUAD had no significant associations with any known pathway.

We noticed that some of the modules of LUAD and LUSC were enriched for the same processes. Nevertheless, a pairwise comparison of each of them suggested that, in most of the cases, the number of overlapping genes in the LUAD and LUSC modules is only a minor fraction. This could suggest that the genes triggering the same pathways had different co-expression signatures in the two cancer types. Pathway-enrichment analyses on DE genes or on the soft-

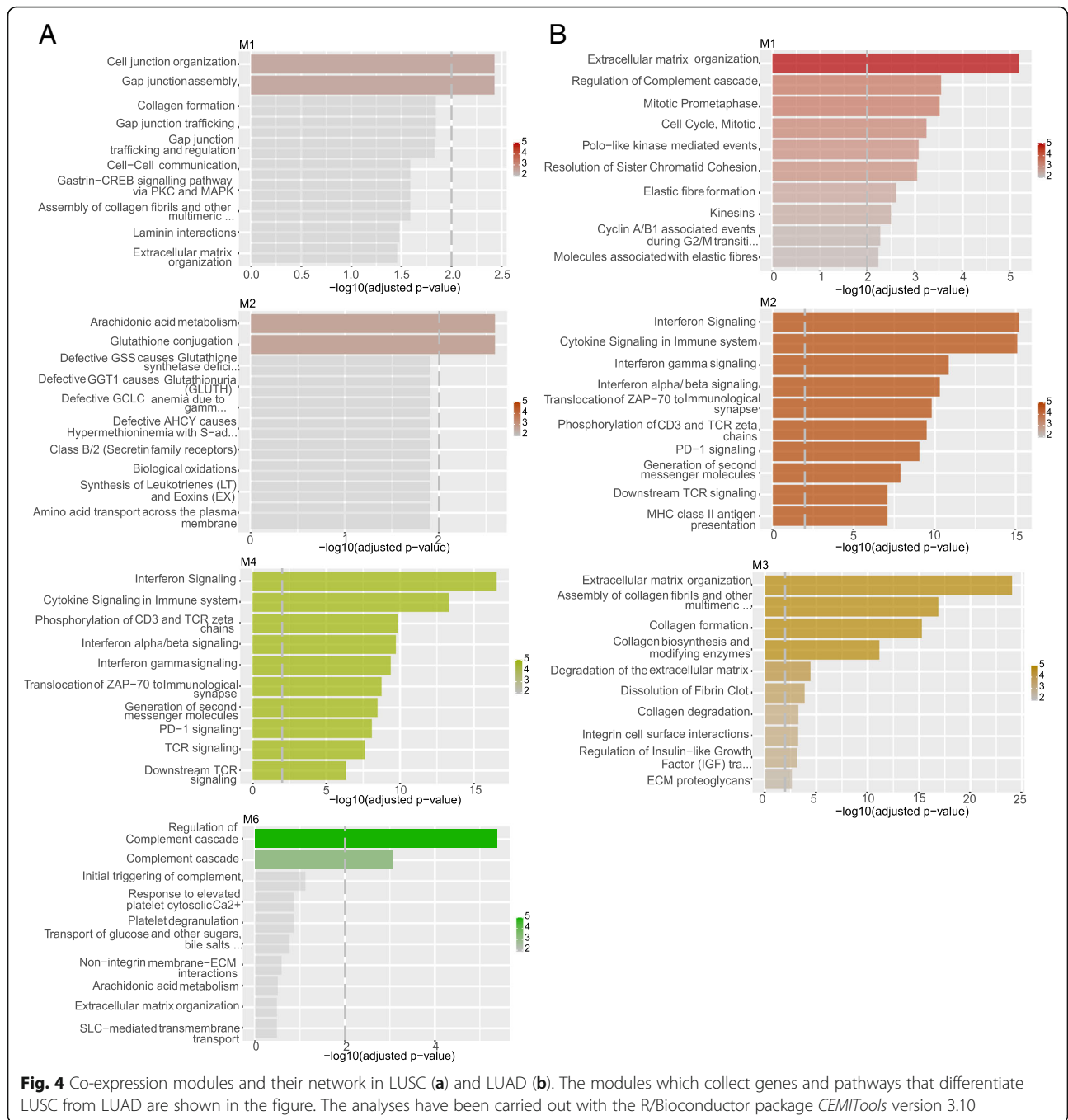
clustering genes also pointed to a down-regulation of proteins involved in the complement cascade (Table 1) and genes related to the immune response in the LUSC samples (Fig. 3d), enforcing the notion of a compromised immune response in LUSC.

Moreover, the M1 and M2 of LUSC were enriched in pathways that have not been found for the LUAD co-expression modules, i.e., pathways related to cellular junctions (M1) and glutathione (M2).

For further analyses, we retained only truly unique genes for LUAD or LUSC within each module. For each module, we extracted the known transcription factors and their targets using the *TRRUST* database as a source of information [81]. We identified a network of transcription factors and their targets for modules 1 and 2 of LUAD, as well as 1, 2,3, and 4 of LUSC (Fig. 5). Out of these, LEF1 was of interest since it can activate NRCAM. The two genes are not only co-expressed in the same LUSC module but also up-regulated in LUSC only. In module 1 of LUSC, we noticed the presence of the up-regulated *CSTA*, which is transcriptionally regulated by the *FOS* transcription factor, along with *TP63* and its target gene *ZNF750*. In module 1 of LUAD, we identified an interesting network between the up-regulated gene *AGR2* and its transcription factor *FOXA1*, along with the activator *SPDEF*.

### Selection of candidate genes and their pathways

We collectively considered the results of the analyses described above with the final goal of proposing a subset of LUAD and LUSC-specific genes for further studies. In particular, we decided to retain only the genes that satisfy the following criteria: i) genes that are up- or down-regulated in a specific cancer type and the opposite in the other cancer type according to the DEA analyses and/or soft-clustering analyses, and ii) genes that belong to the co-expression modules and are genuinely unique for LUAD or LUSC. For each of these genes, we also annotated information on their potential as oncogenes, tumor suppressors or dual role genes; known associations with cancer according to the repository of disease-gene associations from text mining of the literature *DIS-EASES* [82]. Specifically, we verified if they matched with known oncogenes or tumor suppressors through analyses of the *COSMIC* TGs and OCGs collection [77], *TSGDB* [75], *ONGene* [76] or prediction with the *MoonlightR* workflow [37]. For dual role genes, we integrated as a reference for our study the curation from *TSGDB* [75], *COSMIC* [77] and the recently predicted 'double-agent' genes, namely Proto-Oncogenes with Tumor-Suppressor Functions (*POTSF*) [74]. This integrative reference annotation for dual role genes is reported in Additional file 6: Table S2 and in the *GitHub* repository for a total of 152 genes, of which only 14 were all



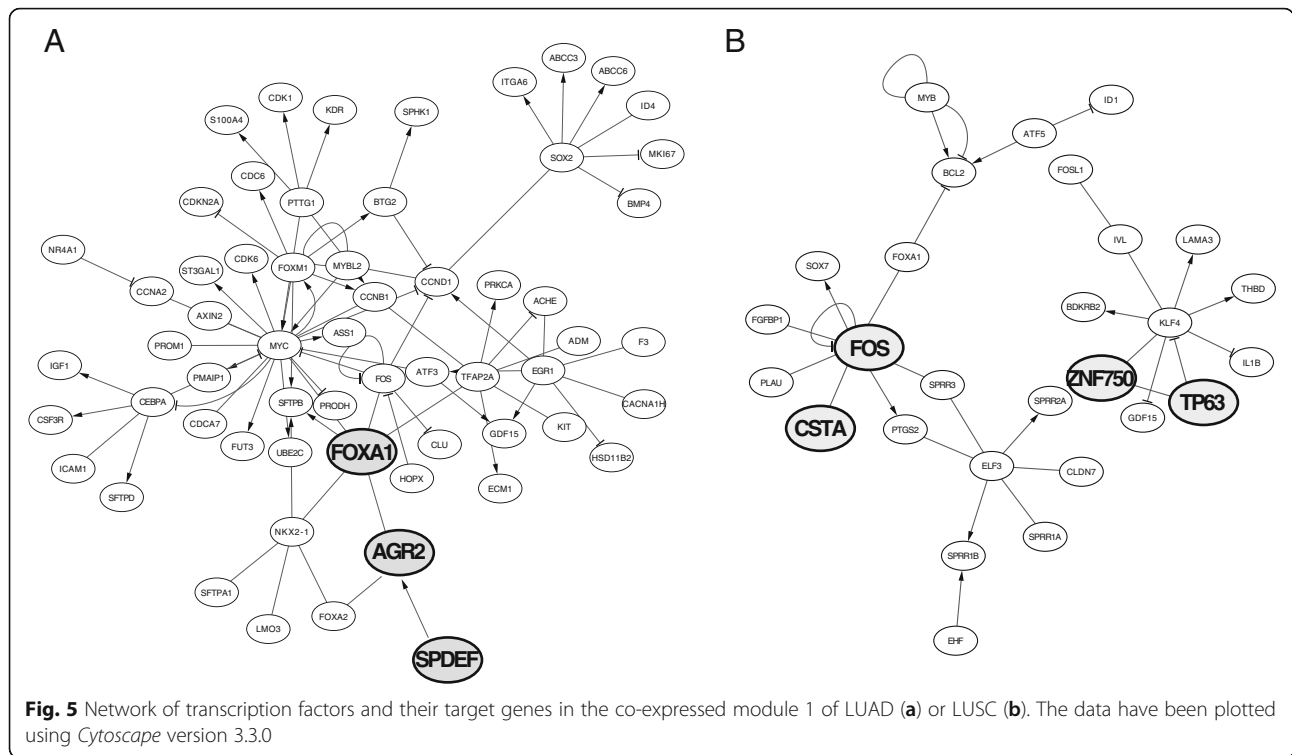
reported in all the three studies. The summary of each of the candidate genes and their annotations is reported in Table 2 and each candidate gene discussed in detail in the Discussion section.

#### Association of the gene signatures with patient survival

Next, we aimed to evaluate if any of the candidate genes had a potential prognostic impact. We carried out survival analyses using a Cox proportional hazard regression with all the candidate genes. We accounted for

different explanatory variables, including the clinical stages, age, and sex of the patients.

We assessed if a difference in the gene expression level (high or low) of the candidate genes could affect the survival rate of the patients. The unique genes with FDR values lower than 0.05 were ITGA6 and FABP5 in LUAD and ICA1 in LUSC (Additional file 7: Table S3). In details, these genes have an FDR associated with the 'group\_low' less than 0.05. In particular, the hazard ratio (the exp. (coef)) is around 0.4 for the three genes. This



means that a patient who has a high level of expression of one of the three genes is 0.4 times as likely to die at any time than a patient with a low level of expression of the same gene. Therefore, the risk associated with high gene expression of ITGA6, FABP5 (in LUAD) or ICA1 (in LUSC) is low, resulting in a better prognosis for these patients.

**Analysis of the candidate genes on independent datasets**

To further strengthen our results, we validated the most interesting markers using two independent datasets, where LUAD and LUSC samples were profiled by transcriptomics techniques with the same experimental setup.

For this analysis, we retained the candidate genes reported in Table 2 for which LUAD’s and LUSC’s upper and lower quartiles were sufficiently separated when compared for the same gene so that they may suggest a potential value as a marker for classification of the two lung cancer types. We then extracted the ones for which gene quantification was available in the validation datasets, and we used unsupervised clustering to verify if they could separate LUAD and LUSC. The results are reported in Fig. 6 and Additional file 4: Figure S3. We could not validate the clustering potential of mucins (except for MUC5B) since the probes for these genes were not available upon the probe set collapse step. MUC5B, HAPBP2, SPDEF, ICA1, FZD7, CHST7, SLC2A9,

ACOX2, KCNK5, ARSE, P2RY1, CSTA, ALDOC, and ANXA8 seem to retain the capability of separating the two lung cancer types. Of note, KCNK5 had been previously proposed in another study [10].

**Discussion**

**Candidate genes in lung cancer and other cancer types**

We performed a literature search to see if any of the candidate genes (Table 2) had been reported in cancer studies. Mucins will be discussed more in detail below. We discussed below the most interesting results.

The hyaluronan binding protein 2 (HABP2) is an extracellular serine protease, which is up-regulated in LUAD and down-regulated in LUSC has been associated with lung cancer [83]. In lung cancer, the up-regulation of hyaluronan in the extracellular matrix regulate the activity of HABP2 and its regulation of cancer progression has been shown in LUAD, in agreement with our results [83]. Our data and the previous findings suggest that HABP2 may be a valuable target to study for diagnostic and therapeutic purposes.

KCNK5 is a two-pore potassium channel and belongs to the K2P family, i.e., a channel which facilitates the extracellular leak of potassium ions [84]. We found this gene to be up-regulated in LUAD and down-regulated in LUSC. Overexpression of members of the K2P family was associated with

**Table 2** Candidate genes to discriminate between LUAD and LUSC in terms of gene expression levels, functions or prognosis

GENE	DEA	Mfuzz cluster	CEMiTool Module	DISEASES	OG	TSG
MUC5B	Up (LUAD) Down (LUSC)	C16(LUAD) C12(LUSC)	M1(LUAD)	None	None	None
HABP2	Up (LUAD) Down (LUSC)	C13(LUAD) C14(LUSC)	M1(LUAD)	3.8	None	None
MUC21	Up (LUAD) Down (LUSC)	C12(LUSC)	M1(LUAD)	4.0	None	None
KCNK5	Up (LUAD) Down (LUSC)	C13(LUAD) C12(LUSC)	M1(LUAD)	None	None	None
ICA1	Up (LUAD) Down (LUSC) <sup>a</sup>	C16(LUAD) C12(LUSC)	M1(LUSC)	1.9	None	None
CSTA	Down (LUAD) Up (LUSC)	C14(LUAD) C11(LUSC)	M1(LUSC)	3.5	None	TSGDB (CST5, CST6)
P2RY1	Down (LUAD) Up (LUSC)	C14(LUAD) C11(LUSC)	M1(LUSC)	2.2	Moonlight (LUSC, P2RY14) COSMIC(P2RY8)	None
ANXA8	Down (LUAD) Up (LUSC)	C15(LUSC)	M1(LUSC)	2.4	None	TSGDB (ANXA1, ANXA7)
FZD7	Up (LUSC)	C14(LUAD) C11(LUSC)	M2(LUSC)	3.7	ONGENE (FZD2) Moonlight (LUSC, FZD4)	None
ITGA6	Up (LUSC) <sup>c</sup>	C14(LUAD) C11(LUSC)	M1(LUAD)	4.2	ONGENE (ITGA3) Moonlight (LUSC, ITGA8)	TSGDB (ITGA5,ITGA7, ITGAV)
CHST7	Up (LUSC) <sup>b</sup>	C14(LUAD) C11(LUSC)	M2(LUSC)	1.6	None	TSGDB (CHST10)
ACOX2	Down (LUSC)	C12(LUSC)	M1(LUAD)	2.2	None	None
ALDOC	Up (LUSC)	C11(LUSC)	M1(LUAD)	3.9	None	None
AQP5	Down (LUSC)	C14(LUSC)	M1(LUAD)	None	None	None
ARSE	Up (LUAD) <sup>a</sup> Down (LUSC)	C12(LUSC)	M1(LUAD)	None	None	None
FABP5	Down (LUAD) Up (LUSC)	C14(LUAD) C15(LUSC)	M1(LUSC)	None	MoonlightL (LUAD,FABP7)	TSGDB (FABP3)
SLC2A9	Down (LUAD) <sup>a</sup>	C14(DOWN) C15(LUSC)	M1(LUSC)	None	None	None
NRCAM	Down (LUAD) <sup>a</sup> Up (LUSC)	C11(LUSC)	M2(LUSC)	2.6	Moonlight (LUSC, OCG)	TSGDB
AGR2	Up (LUAD) Down (LUSC) <sup>b</sup>	C13(LUAD)	M1(LUAD)	4.1	None	None
SPDEF	Up (LUAD) Down (LUSC) <sup>b</sup>	C11(LUAD) C12(LUSC)	M1(LUAD)	3.6	None	None

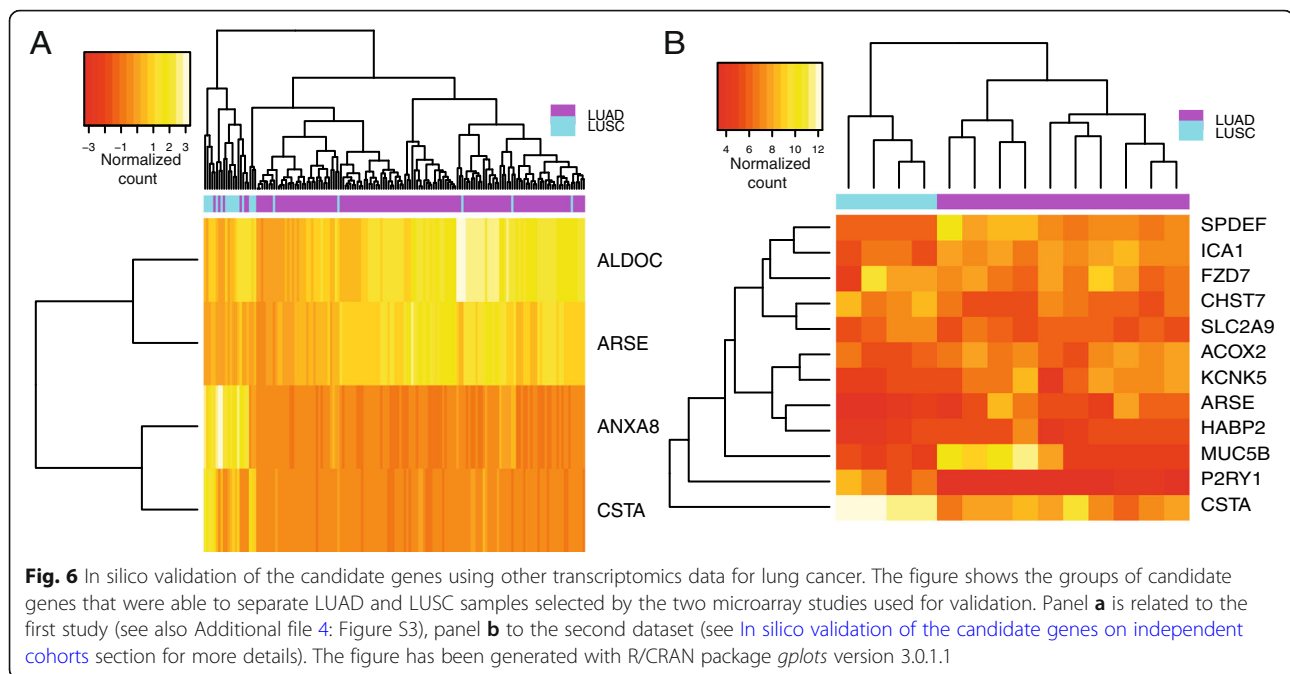
<sup>a</sup>‘N.S.’ indicates not significant results. The *DISEASES* Z-score are provided in the table. <sup>a</sup>, <sup>b</sup>, and <sup>c</sup> indicate a significant DE when the unified recount, the TCGA pair dataset or both are used. We did not find any predicted dual-role genes for any of the candidate genes

different cancer types [84] with some exceptions, such as KCNK4 which was down-regulated. Our results suggest that KCNK5 may be used to classify the two lung cancer types under investigation and that this gene is not necessarily only down-regulated in cancer.

Cystatins, such as CSTA, are cysteine protease inhibitors that regulate different physiological processes [85]. Proteins of this superfamily are classified as tumor suppressors by TSGDB. CSTA is down-regulated in LUAD and up-regulated in LUSC, in our study. CSTA deregulation has been associated with different cancer types [85], and specifically

breast cancer. Breast tumors with positive CSTA expression are associated with poor patient outcome [85]. The study on the potential of CSTA as a prognostic marker in LUSC deserves further investigation, along with its regulation by the FOS transcription factor (Fig. 5).

The purinergic receptor P2Y1 (P2RY1) was down-regulated in LUAD but up-regulated in LUSC, according to our study and also identified as a possible LUSC oncogene by *Moonlight*. Of note, its levels can be regulated by the miR-34b-3p microRNA in bladder cancer [86]. Low levels of P2RY1 contribute to the repression of chemoresistance in a concerted



action with *CCND2* [86] and, in light of our results, it may be an interesting target in LUSC.

Annexin A8 (*ANXA8*) is down-regulated in LUAD and up-regulated in LUSC, according to our study has been classified as tumor suppressors in TSGDB, suggesting that a similar role in LUAD would benefit of further investigation. At the best of our knowledge, the role of *ANXA8* in calcium fluctuation-mediated HIF-1 $\alpha$  transcriptional activation and cell viability has been studied only in pancreatic cancer [87].

Frizzled-7 (*FZD7*) is up-regulated in LUSC, and it is a protein associated with the WNT signaling pathway [88] - a usual suspect in cancer. WNT proteins are secreted glycoproteins, which bind an extracellular cysteine-rich domain of the Frizzled receptor family. *FZD7* has been showed as up-regulated in a variety of cancer types including colorectal cancer, hepatocellular carcinoma, and certain breast cancer subtypes [89]. Our data link *FZD7* to LUSC. Other genes of the family are classified as possible oncogenes in our study, and the up-regulation of *FZD7* would be worth further studies in LUSC since *FZD7* is a known pharmacological target. Small peptides or molecules have been reported to inhibit its activity and, as a consequence, suppress the  $\beta$ -catenin-dependent tumor growth [89].

Integrin alpha 6 (*ITGA6*) is also up-regulated in LUSC in our study and other members of the *ITGA* family classified either as oncogenes or tumor suppressors (Table 2). Integrins mediate interactions with the extracellular matrix but also drive intracellular

communication from the tumor microenvironment leading to migration and invasion. In this context, *ITGA6* has been linked to cancer stemness and invasiveness in breast cancer through a HIF-dependent mechanism [90]. Its HIF-dependent up-regulation could be worth exploring in LUSC as well, especially in connection to the link between stemness and resistance to cancer therapy [91].

*AQP5* is down-regulated in LUSC and is an aquaporin protein, i.e., a water channel [92]. *AQP5* has been reported with a role in invasion in lung cancer, an effect mediated at the protein level and connected with its phosphorylation [92]. The fact that we observed it down-regulated in LUSC, thus, does not imply that its protein level and the protein activity will be affected in this lung cancer type.

*FABP5* is a fatty acid-binding protein, which we found down-regulated in LUAD and up-regulated in LUSC. Other *FABP* proteins have been classified as tumor suppressors or oncogenes, according to our analyses (Table 2), suggesting a complex context-dependent role in cancer. *FAB5* has been linked with tumor cell growth, metastasis, and, in certain cases, poor prognosis in other cancer types [93–95]. Therefore, it would be a valuable future direction to explore its role as a marker discriminating LUAD and LUSC in lung cancer tissues.

*NRCAM* is a neuron-glia-related cell adhesion molecule which is mostly expressed in neurons. Recently, it was also linked to other tissues and cancer types, such as lung adenocarcinoma [96] or thyroid [97]. Our findings are consistent with its down-regulation in LUAD, due to overexpression of *CDH2* [96]. On the other hand, *NRCAM* is

clearly up-regulated in LUSC, entailing a promising marker to discriminate the two lung cancer types.

Another interesting candidate is AGR2, which has been associated with lung cancer, as a prognostic marker [98–100]. Our results point to an up-regulation of AGR2 in LUAD, consistently with the findings in literature and down-regulation in LUSC.

#### ***Genes involved in O-glycosylation of mucins are differentially regulated in different lung cancer types***

Our analyses pinpointed a differential regulation of different genes involved in the O-glycosylation of mucins. These genes are up-regulated in LUAD and down-regulated in LUSC.

Mucins are heavily glycosylated proteins where glycosylation is relevant to their function. Under normal conditions, mucins serve as a protective barrier for epithelial lung cells [101]. When dysregulated, these proteins promote cancer progression and metastasis [102]. During cancer progression, mucins can alone or in combination with different tyrosine kinase receptors mediate cell signals for growth and survival of cancer cells. Expression of certain mucins, such as MUC1 or MUC4, have been associated with lung cancer in other studies, and associated with poor prognosis for some patients [102]. Due to this key role in oncogenesis, mucins are emerging as attractive targets for novel therapeutic approaches to treat lung cancer [102].

Our results suggest that both membrane-bound (e.g., MUC21) and secreted mucins (e.g., MUC5B) contribute to the differences between LUAD and LUSC.

Tumors which overexpress MUC5B have been linked to tendencies for relapse and/or metastasize postoperatively in comparison to non-expressing tumors [103]. This finding suggests that LUAD patients could suffer from these events more often than the ones with a LUSC subtype. MUC5B, which we found up-regulated in LUAD and down-regulated in LUSC (Table 2), has also been associated with an aggressive profile in breast cancer. This gene could be targeted to slow down tumor growth and metastasis [104]. Moreover, MUC5B silencing was shown to reduce chemo-resistance of breast tumor cells [105], suggesting this as an interesting target also for LUAD, where we found MUC5B as one of the up-regulated genes. Mucin expression and its link to chemotherapy resistance has been reported even more broadly in cancer [106].

Mucins are amenable drug targets, as attested by MUC1 which can be targeted by immunotherapy thanks to the availability of T-cell specific antigenic epitopes. Vaccines have been proposed, along with

aptamer-based drugs (for a review [102]). Despite several studies on mucins in lung cancer, these have only scraped the surface of a complex and intricate interplay where also the interactions between the different mucins can add an extra level of undisclosed complexity. Our results suggest that more studies focusing specifically on MUC5B and MUC21 are needed. The opposite behavior of these two genes in LUAD and LUSC and the overexpression in LUAD suggest the possibility of exploiting them (or the enzymes regulating mucin glycosylation) as drug targets for LUAD-specific therapy.

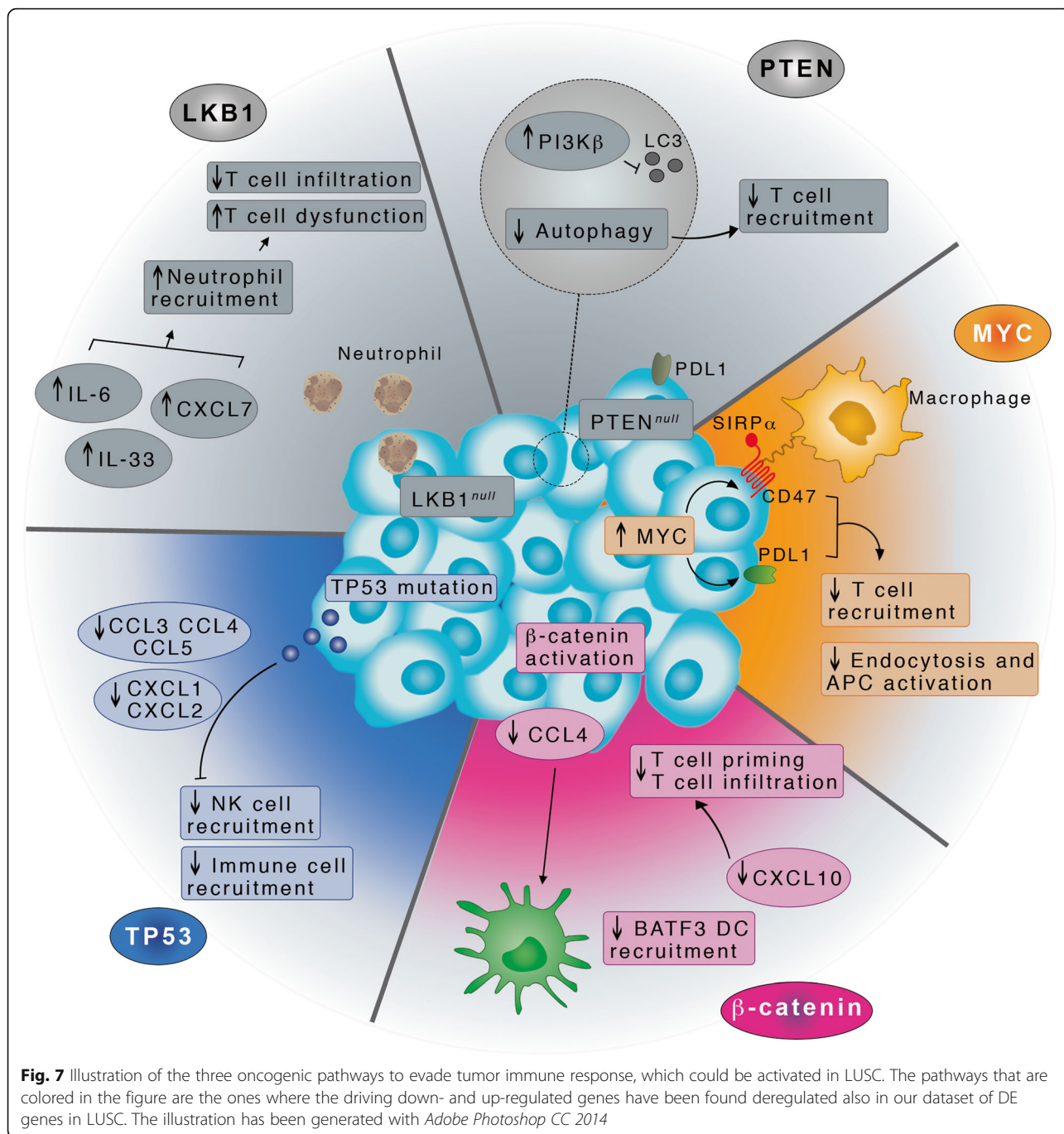
#### ***LUSC and the activation of oncogenic pathways for evasion of antitumor activity***

Generally, an enhanced immune response in cancer can be exploited for therapeutic purposes [107]. We here observed that LUSC seems to be immune-compromised with a signature of down-regulation of the complement cascade and other key genes for immune response. Our results nicely fit within the overall difference in tumor immune landscape in LUAD and LUSC [108]. To further verify that the differences do not come from differences in the composition of the tumor microenvironment between the two cancer types, we also carried out an exhaustive set of differential expression analyses, correcting for the population of the different cellular infiltrations.

Recently, five main oncogenic pathways have been reported [109] that are associated with the evasion of antitumor immunotherapy. The activation of these pathways relies on the dysregulation or mutations of usual suspects in cancer such as p53, cMYC, and the  $\beta$ -catenin/WNT. These genes are the upstream regulators of the evasion pathways and act through a well-orchestrated cascade of other more specific deregulated set of genes (Fig. 7). The oncogenic pathways for evasion of immune response in tumor cells have the ultimate effect of impairing the induction or execution of a local antitumor immune response, which also explains the resistance to certain therapies. We compared the driver up- or down-regulated genes associated to each of these oncogenic pathways [109] with the up- and down-regulated genes in LUSC found in our study. We observed that three evasion pathways (colored in Fig. 7) would be the most suitable candidates to explore further for LUSC, in the direction of the design of new tailored therapies.

#### **Conclusions**

This study allowed to shed new light on the differences between two lung cancer types, i.e., LUAD and LUSC. In addition, we here provided a solid



biostatistics and bioinformatics framework for the interpretation of gene expression data. Our data highlight the importance of a careful assessment of protocols for differential expression analyses since too simplistic approaches without the proper information in the design matrix can result in discordant signatures.

We predicted two potential dual role genes (IL6 and KRT23) in LUAD and LUSC. Our analyses also showed

that LUAD and LUSC differentiate for the biological processes that are altered. Specifically, LUAD features an up-regulation of genes involved in the O-linked glycosylation of mucins, where MUC5B and MUC21 has the potential for target therapy against LUAD. On the other hand, LUSC seems to be associated with a down-regulation of the complement cascade genes and more generally the innate immune response. These events might be triggered, in LUSC, by the activation of three key oncogenic pathways,



stimulated by p53, cMYC and  $\beta$ -catenin that impair the induction of execution of a local antitumor immune response. Future studies on the role of these pathways in LUSC may provide interesting opportunities for drug treatments tailored to this challenging lung cancer type.

We also identified and validated in silico a gene set that could be explored to classify LUAD and LUSC in cancer patient samples. Some of the candidate genes and pathways identified in our study are usual suspects in lung cancer or other cancer types, attesting the validity of our approaches. Moreover, other candidate genes have been poorly investigated, and they could entail novel mechanisms in LUAD and LUSC, deserving attention in future investigations.

## Additional files

**Additional file 1: Text S1** Comparison of different curations of the datasets. The supplementary file includes a detailed comparison of the results achieved using different curations of the TCGA LUAD and LUSC datasets. (DOCX 17 kb)

**Additional file 2: Figure S1** Analysis of the 820 up-regulated genes identified only by *edgeR-TCGAb*. The supplementary figure includes results about the changes in gene expression of the group of up-regulated genes identified only with the old implementation of DEA in *TCGAbio-links*. (DOCX 1634 kb)

**Additional file 3: Figure S2** GO-enrichment analyses. We reported an example of the results of GO-enrichment analyses for the up-regulated genes in LUAD. (DOCX 391 kb)

**Additional file 4: Figure S3** Additional plot for in silico independent validation of the candidate genes. The heatmap includes the full list of candidate genes using the data from the first validation dataset. See Section 2.10 for more details. (PDF 24 kb)

**Additional file 5: Table S1** Summary of DEGs that have been detected by the different methods or using different dataset curations. The table reports information on the number of up- and down-regulated genes found for each DEA in which we used either a different DEA method or a different curation of the tumor samples. (DOCX 13 kb)

**Additional file 6: Table S2** Integrative annotation for dual role genes. A curation from available datasets of dual role genes. (TXT 2 kb)

**Additional file 7: Table S3** Survival analysis. The table contains all the data from the cox regression on the candidate genes. (XLSX 14 kb)

## Acknowledgments

The authors would like to thank Antonio Colaprico (University of Miami), Francesco Russo (NNF Center for Protein Research, University of Copenhagen) and Vanna Albieri (Danish Cancer Society Research Center) for fruitful discussion and comments. We also would like to thank Matteo Lambrughini for assistance with the illustrations for Fig. 7. The results shown here are in whole or part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>

## Ethics approval and consent to participate

Not applicable.

## Authors' contributions

Conceptualization: EP; Data curation: ML, EP; Formal Analysis: ML, IP, EP; Funding Acquisition: EP; Investigation: ML, EP; Methodology: ML, TT, MV, MM, IP, EP; Project Administration: EP; Resources: EP; Software: ML, IP, MM, EP; Supervision: EP; Validation: EP, ML; Visualization: ML, IP, EP; Writing-Original Draft Preparation: EP; Writing-Review and Editing: EP with inputs from all the coauthors. All authors have read and approved the final submitted manuscript.

## Funding

The project was supported by a KBVU Pre-graduate scholarship 2017 to M.L. in EP group and the Innovation Fund Denmark grant 5189-00052B to EP. EP group is also a part of the Center of Excellence in Autophagy, Recycling and Disease (CARD) funded by the Danish National Research Foundation (DNRF125). The funding bodies did not have any role in the design of the study and collection, analysis and interpretation of data and in writing of the manuscript.

## Availability of data and materials

The datasets generated and/or analyzed during the current study are available in our Github repository: [https://github.com/ELELAB/LUAD\\_LUSC\\_TCGA\\_comparison](https://github.com/ELELAB/LUAD_LUSC_TCGA_comparison)

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests

Received: 18 September 2018 Accepted: 22 July 2019

Published online: 20 August 2019

## References

- Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification. *J Thorac Oncol*. 2015;10:1243–60. Available from: <https://doi.org/10.1097/JTO.0000000000000630>.
- Kadota K, Sima CS, Arcila ME, Hedvat C, Mark KG, Jones DR, et al. KRAS mutation is a significant prognostic factor in early stage lung adenocarcinoma. *Am J Surg Pathol*. 2016;40:1579–90.
- Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong K-K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat rev Cancer*. *Nat Publ Group*. 2014;14:535–46. Available from: <https://doi.org/10.1038/nrc3775>.
- Kadota K, Yeh Y-C, D'Angelo SP, Moreira AL, Kuk D, Sima CS, et al. Associations between mutations and histologic patterns of mucin in lung adenocarcinoma: invasive mucinous pattern and extracellular mucin are associated with KRAS mutation. *Am J Surg Pathol*. 2014;38:1118–27.
- Shea M, Costa DB, Rangachari D. Management of advanced non-small cell lung cancers with known mutations or rearrangements: latest evidence and treatment approaches. *Ther Adv Respir Dis*. 2016;10:113–29.
- Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, Dowlati A, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med*. 2006;355:2542–50. Available from: <https://doi.org/10.1056/NEJMoa061884>.
- Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*. 2002;62:4963–7.
- Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, et al. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer*. 2011;129:355–64.
- Navab R, Strumpf D, Bandarchi B, Zhu C, Pintilie M, Rohan V. Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. *Proc Natl Acad Sci U S A*. 2011;108:7160–5.
- Girard L, Rodriguez-Canales J, Behrens C, Thompson DM, Botros IW, Tang H, et al. An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clin Cancer Res*. 2016;22:4880–9.
- Cui R, Meng W, Sun H-L, Kim T, Ye Z, Fassan M, et al. MicroRNA-224 promotes tumor progression in nonsmall cell lung cancer. *Proc Natl Acad Sci U S A*. 2015;E4288–97. Available from: <https://doi.org/10.1073/pnas.1502068112>.
- Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res*. 2006;66:7466–72.
- Hamamoto J, Soejima K, Yoda S, Naoki K, Nakayama S, Satomi R, et al. Identification of microRNAs differentially expressed between lung squamous cell carcinoma and lung adenocarcinoma. *Mol Med Rep*. 2013;8:456–62.
- Liu J, Yang XY, Shi WJ. Identifying differentially expressed genes and pathways in two types of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. *Genet Mol Res*. 2014;13:95–102 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24446291>.

15. Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*. 2018;19:1–13.
16. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
17. Gonzalez-Valbuena E-E, Treviño V. Metrics to estimate differential co-expression networks. *BioData Min*. 2017;10:32.
18. Wolf DM, Lenburg ME, Yau C, Boudreau A, Van't Veer LJ. Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One*. 2014;9:e88309.
19. Gov E, Arga KY. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Sci Rep*. 2017;7:1–10.
20. Wang W, Hu B, Wang X, Chen J, Qian X, He Y. Candidate genes in gastric cancer identified by constructing a weighted gene co-expression network. *PeerJ*. 2018;6:e4692.
21. Shi Z, Derow CK, Zhang B. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst Biol*. 2010;4:74.
22. Han L, Hei N, Li J, Yuan Y, Liang H, Yang Y. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun* [internet]. *Nat Publ Group*. 2014;5:1–9. Available from: <https://doi.org/10.1038/ncomms4231>.
23. Federoff HJ, Meehan RR, Villoslada P, Baranzini S, Chung KF, Sterk PJ, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med*. 2014;6:1–11.
24. Belling K, Rajpert-De Meyts E, Dalgaard MD, Jensen AB, Skakkebaek NE, Brunak S, et al. Klinefelter syndrome comorbidities linked to increased X chromosome gene dosage and altered protein interactome activity. *Hum Mol Genet*. 2017;26:1219–29.
25. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer genome Atlas pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24071849>.
26. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer genome Atlas (TCGA): an immeasurable source of knowledge. *Wspolczesna Onkol*. 2015;11A:68–77.
27. Tian F, Zhao J, Fan X, Kang Z. Weighted gene co-expression network analysis in identification of metastasis-related genes of lung squamous cell carcinoma based on the Cancer genome Atlas database. *J Thorac Dis*. 2017;9:42–53.
28. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract>.
29. Cancer T, Atlas G, Collisson EA, Campbell JD, Brooks AN, Berger AH, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543–50 Available from: <http://www.nature.com/doi/10.1038/nature13385%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25079552%5Cnhttp://dx.doi.org/10.1038/nature13385>.
30. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519–25 Available from: <http://www.nature.com/doi/10.1038/nature11404>.
31. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 2017;35:319–21.
32. Carithers LJ, Moore HM. The genotype-tissue expression (GTEx) project. *Biopreserv Biobank*. 2015;13:307–8. Available from: <https://doi.org/10.1089/bio.2015.29031.hmm>.
33. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. *Nature* [internet]. *Nat Publ Group*. 2013;502:333–9 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3927368&tool=pmcentrez&rendertype=abstract>.
34. Goldman M, Craft B, Swatloski T, Elliott K, Cline M, Diekhans M, et al. The UCSC cancer genomics browser: update 2013. *Nucleic Acids Res*. 2013;41:951–9.
35. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*. 2014;158:929–44. Available from: <https://doi.org/10.1016/j.cell.2014.06.049>.
36. Jia P, Pao W, Zhao Z. Patterns and processes of somatic mutations in nine major cancers. *BMC Med Genomics*. 2014;7:11 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3942057&tool=pmcentrez&rendertype=abstract>.
37. Colaprico A, Olsen C, Cava C, Terkelsen T, Silva TC, Olsen A, et al. Moonlight: a tool for biological interpretation and driver genes discovery. *bioRxiv*. 2018. Article number: 265322. <https://doi.org/10.1101/265322>.
38. Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, Piccolo SR. Alternative preprocessing of RNA-sequencing data in the Cancer genome Atlas leads to improved analysis results. *Bioinformatics*. 2015;31:3666–72.
39. Cline MS, Craft B, Swatloski T, Goldman M, Ma S, Haussler D, et al. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Sci Rep*. 2013;3:2652 Available from: <http://www.nature.com/srep/2013/131002/srep02652/full/srep02652.html>.
40. Huang Q, Wei H, Wu Z, Li L, Yao L. Preferentially expressed antigen of melanoma prevents lung Cancer metastasis. *PLoS One*. 2016;11:1–15.
41. Hammerman PS, Sos ML, Ramos AH, Xu C, Dutt A, Zhou W, et al. Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. *Cancer Discov*. 2012;1:78–89.
42. Capizzi M, Strappazzon F, Cianfanelli V. MIR3-3HG, a MYC-dependent modulator of cell proliferation, inhibits autophagy by a regulatory loop involving AMBRA1. *Autophagy*. 2017;1:1–41. Available from: <https://doi.org/10.1080/15548627.2016.1269989>.
43. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Carolini D, et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2015;44:gvk1507 Available from: <http://nar.oxfordjournals.org/content/early/2015/12/23/nar.gvk1507.full>.
44. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, et al. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Res*. 2016;5:1542 Available from: <http://f1000research.com/articles/5-1542/v1>.
45. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. 2015;6:8971 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26634437>.
46. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12:115–21.
47. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32:896–902. Available from: <https://doi.org/10.1038/nbt.2931>.
48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26:139–40.
49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47.
50. Lex A, Gehlenborg N, Strobel H. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph*. 2014;20:1983–92.
51. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the TCGAAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLOS Comput Biol*. 2019;15:e1006701 Available from: <http://dx.plos.org/10.1371/journal.pcbi.1006701>.
52. Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J, et al. BioMart: a data federation framework for large collaborative projects. *Database*. 2011;2011:1–15.
53. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*. 2005;21:3439–40.
54. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4:1184–91.
55. Kumar L, Futschik ME. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics*. 2007;23:7 Available from: <http://www.bioinformatics.net/002/000200022007.htm>.
56. Futschik ME, Carlisle B. Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol*. 2005;03:965–88. Available from: <https://doi.org/10.1142/S0219720005001375>.
57. Guangchuang Y, Qing-Yu H. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst*. 2016;12:477–9.
58. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi A J Integr Biol*. 2012;16:284–7. Available from: <https://doi.org/10.1089/omi.2011.0118>.
59. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium*. *Nat Genet*. 2000;25:25–9. Available from: <https://doi.org/10.1038/75556>.

60. Walter W, Sánchez-Cabo F, Ricote M. GOrplot: an R package for visually combining expression data with functional analysis. *Bioinformatics*. 2015;31:2912–4.
61. Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*. 2016;44:D536–41 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4702811&tool=pmcentrez&rendertype=abstract>.
62. Christensen E. Multivariate survival analysis using Cox's regression model. *Hepatology*. 1987;7:1346–58 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3679094>.
63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Royal Stat Soc Ser B*. 1995;57:289–300 Available from: <https://www.jstor.org/stable/2346101>.
64. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*. 2001;98:13790–5. Available from: <https://doi.org/10.1073/pnas.191502998>.
65. Meister M, Belousov A, Xu E, Schnabel P, Warth A, Hoofmann H, et al. Intra-tumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer. *J Bioinforma Res Stud*. 2014;1:1.
66. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res*. 2016;44:W83–89 Available from: <http://nar.oxfordjournals.org/content/early/2016/04/29/nar.gkw199.abstract>.
67. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. 2016;17:218 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27765066>.
68. Sonesson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013;14:91 Available from: <http://www.biomedcentral.com/1471-2105/14/91>.
69. Germain P-L, Vitriolo A, Adamo A, Laise P, Das V, Testa G. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res*. 2016;44(11):5054–67. Available from: <https://doi.org/10.1093/nar/gkw448>.
70. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics*. 2015;14:130–42.
71. Tang M, Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC bioinformatics* [internet]. *BMC Bioinformatics*. 2015:1–14. Available from: <https://doi.org/10.1186/s12859-015-0794-7>.
72. Kao S, Shiau CK, Gu DL, Ho CM, Su WH, Chen CF, et al. IGDB.NSCLC: Integrated genomic database of non-small cell lung cancer. *Nucleic Acids Res*. 2012;40:972–7.
73. Edge SB, Compton CC. The american joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol*. 2010;17:1471–4.
74. Shen L, Shi Q, Wang W. Double agents : genes with both oncogenic and tumor-suppressor functions. *Oncogenesis*. 2018; Available from: <https://doi.org/10.1038/s41389-018-0034-x>.
75. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res*. 2015;44:D0123–D1031.
76. Liu Y, Sun J, Zhao M. ONGene: a literature-based database for human oncogenes. *J Genet Genomics*. 2017;44:119–21. Available from: <https://doi.org/10.1016/j.jgg.2016.12.004>.
77. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43:D805–11. Available from: <https://doi.org/10.1093/nar/gku1075>.
78. Reis ES, Mastellos DC, Ricklin D, Mantovani A, Lambris JD. Complement in cancer: untangling an intricate relationship. *Nat rev Immunol*. *Nat Publ Group*. 2018;18:5–18. Available from: <https://doi.org/10.1038/nri.2017.97>.
79. Espinoza JA, Jabeen S, Batra R, Papaleo E, Haakensen V, Timmermans Wielenga V, et al. Cytokine profiling of tumour interstitial fluid of the breast and its relationship with lymphocyte infiltration and clinicopathological characteristics. *Oncoimmunology*. 2016;5:00 Available from: <https://www.tandfonline.com/doi/full/10.1080/2162402X.2016.1248015>.
80. Su C, Zhou C, Zhou S, Xu J. Serum cytokine levels in patients with advanced non-small cell lung cancer: correlation with treatment response and survival. *Med Oncol*. 2011;28:1453–7.
81. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46:D380–6 Available from: <http://academic.oup.com/nar/article/46/D1/D380/4566018>.
82. Pletscher-Frankild S, Pallegà A, Tsaou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease-gene associations. *Methods*. 2015;74:83–9. Available from: <https://doi.org/10.1016/j.jmeth.2014.11.020>.
83. Lennon FE, Salgia R, Mambetsariev N, Mirzapozazova T, Mambetsariev B, Singleton PA, et al. HAP2 is a novel regulator of Hyaluronan-mediated human lung Cancer progression. *Front Oncol*. 2015;5:1–12.
84. Williams S, Bateman A, O'Kelly I. Altered expression of two-pore domain potassium (K2P) channels in Cancer. *PLoS One*. 2013;8:e74589.
85. Rivenbark AG, Coleman WB. Epigenetic regulation of cystatins in cancer. *Front Biosci*. 2009;14:453–62.
86. Tan Y, Zhang T, Zhou L, Liu S, Liang C. MiR-34b-3p Represses the Multidrug-Chemoresistance of Bladder Cancer Cells by Regulating the CCND2 and P2RY1 Genes. *Med Sci Monit*. 2019;25:1323–35 Available from: <https://www.medscimonit.com/abstract/index/idArt/913746>.
87. Hata H, Tatemichi M, Nakadate T. Involvement of annexin A8 in the properties of pancreatic cancer. *Mol Carcinog*. 2012;53:181–91 Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/mc.21961>.
88. Polakis P. Wnt signaling in Cancer. *Cold Spring Harb Perspect Biol*. 2012;4:1–13.
89. King TD, Zhang W, Suto MJ, Li Y. Frizzled7 as an emerging target for cancer therapy. *Cell Signal*. 2012;24:846–51. Available from: <https://doi.org/10.1016/j.cellsig.2011.12.009>.
90. Brooks DLP, Schwab LP, Krutilina R, Parke DN, Sethuraman A, Hoogewijs D, et al. ITGA6 is directly regulated by hypoxia-inducible factors and enriches for cancer stem cell activity and invasion in metastatic breast cancer models. *Mol Cancer* [internet]. *Mol Cancer*. 2016;15:1–19. Available from: <https://doi.org/10.1186/s12943-016-0510-x>.
91. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018;173:338–354.e15.
92. Gong G, Woo J, Lee J, Kim MS, Jang SJ, Kang SK, et al. Expression of aquaporin 5 (AQP5) promotes tumor invasion in human non small cell lung Cancer. *PLoS One*. 2008;3:e2162.
93. Wang W, Chu H, Liang Y, Huang J, Shang C, Tan H, et al. FABP5 correlates with poor prognosis and promotes tumor cell growth and metastasis in cervical cancer. *Tumor Biol*. 2016;37:14873–83 Available from: <http://link.springer.com/10.1007/s13277-016-5350-1>.
94. Kawaguchi K, Senga S, Kubota C, Kawamura Y, Ke Y, Fujii H. High expression of Fatty Acid-Binding Protein 5 promotes cell growth and metastatic potential of colorectal cancer cells. *FEBS Open Bio*. 2016;6:190–9. Available from: <https://doi.org/10.1002/2211-5463.12031>.
95. Liu R-Z, Graham K, Glubrecht DD, Germain DR, Mackey JR, Godbout R. Association of FABP5 Expression With Poor Survival in Triple-Negative Breast Cancer: Implication for Retinoic Acid Therapy. *Am J Pathol*. 2011;178:997–1008 Available from: <https://www.sciencedirect.com/science/article/pii/S000294401000221X>.
96. Zhuo H, Zhao Y, Cheng X, Xu M, Wang L, Lin L, et al. Tumor endothelial cell-derived cadherin-2 promotes angiogenesis and has prognostic significance for lung adenocarcinoma. *Mol Cancer*. 2019;18:34 Available from: <https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-019-0987-1>.
97. Górka B, Skubis-Zegadło J, Mikula M, Bardadin K, Paliczka E, Czarnocka B. NrCAM, a neuronal system cell-adhesion molecule, is induced in papillary thyroid carcinomas. *Br J Cancer*. 2007;97:531–8 Available from: <http://www.nature.com/articles/6603915>.
98. Hsu Y-L, Hung J-Y, Lee Y-L, Chen F-W, Chang K-F, Chang W-A, et al. Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. *Oncotarget*. 2017;8:104831–54 Available from: <http://www.oncotarget.com/fulltext/21022>.
99. Alavi M, Mah V, Maresh EL, Bagryanova L, Horvath S, Chia D, et al. High expression of AGR2 in lung cancer is predictive of poor survival. *BMC Cancer*. 2015;15:655 Available from: <http://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-1658-2>.
100. Dietel M. Article in Histology and histopathology. 2007 [cited 2019 Mar 22]; Available from: <http://www.hh.um.es>

101. Hollingsworth MA, Swanson BJ. Mucins in cancer: protection and control of the cell surface. *Nat Rev Cancer*. 2004;4:45–60 Available from: <http://www.nature.com/doi/10.1038/nrc1251>.
102. Lakshmanan I, Ponnusamy MP, Macha MA, Haridas D, Majhi PD, Kaur S, et al. Mucins in lung cancer: Diagnostic, prognostic, and therapeutic implications. *J Thorac Oncol*. 2015;10:19–27. Available from: <https://doi.org/10.1097/JTO.0000000000000404>.
103. Yu C-J, Shih J-Y, Lee Y-C, Shun C-T, Yuan A, Yang P-C. Sialyl Lewis antigens: association with MUC5AC protein and correlation with post-operative recurrence of non-small cell lung cancer. *Lung Cancer*. 2005;47:59–67 Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0169500204002594>.
104. Valque H, Gouyer V, Gottrand F, Desseyn JL. MUC5B leads to aggressive behavior of breast Cancer MCF7 cells. *PLoS One*. 2012;7:e46699.
105. Garcia EP, Tiscornia I, Libisch G, Trajtenberg F, Bollati-Fogolin M, Rodriguez E, et al. MUC5B silencing reduces chemo-resistance of MCF-7 breast tumor cells and impairs maturation of dendritic cells. *Int J Oncol*. 2016;48:2113–23.
106. Jonckheere N, Skrypek N, Van Seuningen I. Mucins and tumor resistance to chemotherapeutic drugs. *Biochim Biophys Acta*. 2014;1846:142–51. Available from: <https://doi.org/10.1016/j.bbcan.2014.04.008>.
107. Guo L, Zhang H, Chen B. Nivolumab as programmed Death-1 (PD-1) inhibitor for targeted immunotherapy in tumor. *J Cancer*. 2017;8:410–6.
108. Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M. Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. *J Thorac Oncol*. 2017;12:943–53. Available from: <https://doi.org/10.1016/j.jtho.2017.03.010>.
109. Spranger S, Gajewski TF. Impact of oncogenic pathways on evasion of antitumour immune responses. *Nat Rev Cancer*. 2018;18:139–47. Available from: <https://doi.org/10.1038/nrc.2017.117>.

### Publisher's Note

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

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

