Systematical identifications of prognostic meaningful lung adenocarcinoma subtypes and the underlying mutational and expressional characters

Ting Lei (lyu17709872529@163.com)
Dalian Medical University

Zhenyang Lv
Dalian medical universty

Research article

Keywords: lung adenocarcinoma, prognosis, causal regulatory network, prognosis master regulator, subtypes

Posted Date: August 28th, 2019

DOI: https://doi.org/10.21203/rs.2.12424/v1

Version of Record: A version of this preprint was published on January 27th, 2020. See the published version at https://doi.org/10.1186/s12885-019-6462-y.

Abstract

Lung adenocarcinoma (LUAD) is one of the most common cancer types. However, the high heterogeneity and complexity of LUAD hinder effective treatments. This study aimed to identify the key prognosis impacting genes and the corresponding subtypes for LUAD. Specifically, the cox proportional hazards model was combined with a causal regulatory network to help reveal which genes play master roles among numerous prognosis impacting genes, and sub-types were identified based on expressional profiles of the master genes. As results, a collection of 75 genes were recognized as the master prognosis impacting genes, where some were enriched in mTOR signaling and lysosome pathways. Based on their expressions, the LUAD patients were separated into two sub-types displaying remarkable differences in expressional profiles, prognostic outcomes and genomic mutations. Meanwhile, the two subtypes were re-discovered from two additional LUAD cohorts based on only the top-10 important master genes. This study provides a comprehensive description on the key prognosis-relevant genes and an alternative way to classify LUAD subtypes which can promote LUAD precision treatment.

Introduction

Lung cancer is one of the most frequent malignant neoplasm, and one of the major causes of cancer death among both males and females around the world (1, 2). Lung cancer is a highly heterogeneous and complex disease which includes many subtypes. Histologically, lung adenocarcinoma (LUAD) is the most common one. Recently, molecular targeted therapies have improved the treatments for LUAD, particularly for patients with specific mutations in *EGFR* (3), *ALK* (4, 5), *RET* (6) and *ROS1*(7). Meanwhile, promising novel targets like *KRAS* (8) and *MET* (9) are being studied . However, the high heterogeneity and complicated molecular patterns of LUAD limit the benefits of these targeted therapies to only specific patients, leaving large amount of LUAD patients without effective therapeutic drugs. It is essential to obtain a more comprehensive view on the molecular mechanism of LUAD, rather than solely focusing on the therapeutically targeted mutations.

Owing to the advantage of high-throughput omics technology, large scale descriptions on the molecular characters of LUAD have been achieved (10). Accordingly, the potential complicated molecular mechanism underlying LUAD has been more extensively explored by mining the LUAD relevant omics data (11) (12). These omics based studies help identified a series of prognosis or diagnosis relevant biomarkers which can provide novel and promising treatment targets. However, the omics-based cancer investigations, which mainly depend on mutation significance examination, differential analysis, or expression-based survival analysis, will generate a larger number of interesting items, either in gene or protein level (10). It is unquestionable these genome- or proteome-wise identifications generate certain mechanical or clinical meaningful biomarkers (13, 14). However, human body is a complex organism, these interesting items must function in a collective way rather than individually. A big challenge is how to understand the mutual

associations among these most functional items and recognize the most functional multi-item sets from the interesting items. Besides, the consistency of the identified molecular patterns across different datasets is also an important issue.

To solve these issues, we put-forward a causal network based framework to help systematically investigate on the prognosis-relevant genes and their mutual association patterns underlying LUAD. Through this study, a causal regulatory network among prognosis relevant genes will be constructed. Based on this network, we can identify the master prognosis impacting genes and the prognosis-meaningful LUAD subtypes can be recognized.

Materials And Methods

TCGA Data

The mutation and RNA-seq data for LUAD were obtained from TCGA (15). Firstly, we downloaded both kinds of data for 33 types of cancers from The NCI's Genomic Data Commons (GDC) (https://gdc.cancer.gov/about-data/publications/pancanatlas). The mutational data were saved in mutation annotation format (16), and the RNA-seq data were saved in a tab file. The maf data was processed by the R package maftools, and the RNA-seq data were preprocessed based on the voom algorithm (17) in the R package limma (18). For this study, we extracted the data corresponding to LUAD patients.

Pathway data

Pathway information were integrated from two databases including KEGG and Molecular Signatures Database (MsigDB, http://software.broadinstitute.org/gsea/msigdb) (19), and pathway names as well as genes belonging to each pathway were extracted from the databases.

Identification of significant SMGs

MutSigCV(version 1.3.4) (20) was applied on the the maf mutation file to recognize significant SMGs where the significance threshold was set as q-value <0.1. Then, we utilized the maftools to visualize the mutation information of these significant SMGs among TCGA LUAD patients.

Survival analysis based on gene expressions

The clinical information of all TCGA-LUAD patients was also obtained from the GDC. For the mRNA expression data, we removed genes with more than 70% of zero values, and analyzed the prognosis impacts for the remained genes. For each such gene, we utilized a Cox proportional hazards (Coxph) regression model in the R package “survival” (21) to examine whether the expression level of this gene has a significant influence on the survival rate. According to the Coxph results, genes with p-values less than 0.05 were regarded as prognosis-relevant, and if the regression coefficients are larger than 0, then higher expression levels will correspond to worse survival rates, otherwise, higher expression levels will correspond to better survival rates.

Identification of master prognosis impacting genes by a causal regulatory network

According to the Coxph-based survival analysis results, genes with p-values less than 0.01 and absolute coefficient values larger than 0.2 were taken as the prognosis relevant genes. Then, based on the mRNA expression profiles of these genes, we computed the bi-weight mid-correlations (22) among all pair-wise genes. To recognize the most likely causal correlations, we further estimated a causal regulatory network based on the correlation matrix. This causal regulatory network was a directed acyclic graph describing the conditional independence relationships. It was estimated by the PC-algorithm in R package pcalg (23). Since this causal regulatory network was directed, a summarized node degree was calculated as the number of all out-going edges minus the number of all in-coming edges (i.e., out degree - in degree). Then nodes were taken as the master prognosis impacting genes if the absolute values of their summarized degrees were larger than certain threshold.

Unsupervised clustering of patients based on SMG relevant genes and pathways

We clustered the LUAD patients into two groups based on the mRNA-level expression matrix of the master prognosis regulating genes. This expression matrix was scaled by subtracting the mean level and being divided by the standard derivation with respect to each individual gene. Based on the scaled expression matrix, we applied a consensus clustering method in the R package “ConsensusClusterPlus” (24) to cluster the patients into 2 clusters where “ partitioning around medoids” was chosen as the basic clustering algorithm.

To further examine the significance of the prognosis effects generated by the two clusters, we also repeated the clustering processes 100 times, where each time we clustered the LUAD based on the expression profiles of randomly-selected genes (with the same number of the prognosis-relevant genes). Thus, we obtained the distribution of the log-rank p values for randomly conditions.

Evaluate the importance of genes for the clustering results

After clustering the patients into 2 clusters, we used the random forest (RF) (25) algorithm to evaluate the importance of all genes in the input expression matrix for predicting the accurate cluster labels. These genes were ranked by the importance score. Besides, we also examined enrichment significance of the top-50 important genes in each pathway based on the hypergeometric distribution.

Estimate the relevance between the prognosis relevant genes and SMGs

For each SMG, we separated the samples into mutated and wild type sets, and utilized T-test (un-paired, two-sides) to identify which genes were differentially expressed between mutated and wild type set in the transcriptomics data, then genes with FDR less than 0.1 were regarded as the SMG relevant.

Validating the LUAD subtypes based on the other independent lung cancer cohorts

The mRNA expression matrix and the corresponding clinical information for two lung cancer cohorts (GSE30219 and GSE31210) were downloaded from GEO database by the R package 'GEOquery' (26). We constructed a cluster-label predictor based on the TCGA-LUAD expression matrix of the top-10 important genes. This predictor was trained by a generalized linear model (implemented by the R package "glmnet" (27)). We used this predictor on the GEO datasets to predict the corresponding cluster-labels for each patient, and survival differences of the two predicted clusters were tested by Coxph model and the corresponding survival curves were estimated by the Kaplan-Meier method.

Drug and gene interactions

All drug and gene interaction information were obtained from the DGIdb (28) which included both the known and reported drug-gene interactions.

Statistical analysis

The statistical analysis and relevant computations were implemented by R. Detailed information was described in corresponding method sections.

Results

Significant somatic mutations in LUAD

According to the gene mutation data of 515 LUAD patients in The Cancer Genome Atlas (TCGA) (29), we identified 20 significant somatic mutated genes (SMGs) by MutSigCV (q-value < 0.1, Figure 1). Most of these SMGs like *TP53*, *KRAS*, *KEAP1*, *STK11*, *EGFR*, *NF1*, *BRAF*, *SMARCA4*, etc. have already been identified in the other studies, especially the study

conducted on the previously collected 230 TCGA-LUAD patients (10). The significant mutations of *COL11A1* (q-value = 8.8e-06, mutation rate = 21%) is rarely identified in the previous LUAD studies. However, expressions of *COL11A1* have shown associations with prognostic factors, pathological stages, and lymph node metastasis in non-small cell lung cancer (NSCLC) in some previous studies (30-33). The mutations of *COL11A1* might also be one of the driven factors for LUAD.

To evaluate the most direct mutational effects, we compared the expressions of these SMGs in the mutated and wild type samples. We found that some of the SMGs can lead to significant expressional alterations. For instance, the mutations of *RBM10* will significantly reduce its expressions while mutations of *EGFR* might improve the expression levels (Figure S1A). However, only three of the expressional alterations (*SMAD4*, *CDKN2A* and *TCEAL5*) can lead to significant prognosis influence (Figure S1B).

Genome-wide identification of prognosis relevant genes for LUAD

The genomic mutations can not only influence the functions of the mutated genes, but may also generate remarkable effects on down-stream cascades, thus leading to the final impacts on clinical phenotypes, like the prognostic outcomes. To gain a comprehensive understanding on the prognosis impacts, we attempted to get a genome-wide estimation of the prognosis relevant genes for LUAD based on a Cox proportional hazards (Coxph) model (see Materials and methods). Through this analysis, we observed that a great number of genes were associated with LUAD prognosis (Figure 2A). The high expression levels of some genes like *USP4*, *DTNBP1* may contribute to better survival rates (these genes were termed as favourable ones), while some of the others (termed as un-favourable ones) such as *AHSA1* and *MESDC2* may lead to worse survivals (Figure 2A and 2B). AHSA1 is a co-chaperone of HSP90AA1, and a previous study has revealed that it is involved in the proliferation, migration, and invasion processes of osteosarcoma (34). Here, we observed that higher expression of AHSA1 was associated with a worse survival rate (Figure 2B). USP4 is a deubiquitinating enzyme which may inhibit p53 by deubiquitinating an important p53 ubiquitin ligase ARF-BP1 (35). Correspondingly, many studies have identified the oncogene effects of USP4 (36). Here, on the contrary, we found that higher expressions of USP4 in LUAD may lead to better prognosis. This may be caused by the highly heterogeneity of cancers and the alternative deubiquitinating substrates for USP4. Consequently, the prognosis effects of these identified genes may be conditional.

Meanwhile, although more favourable genes were observed than un-favourable ones, both kinds of genes exhibited a large exploring space (Figure 2C). This indicates the complicated molecular patterns underlying LUAD pathology and implies the importance to

identify the most meaningful genes which may play the master roles in regulating the disease progression.

Construction of a potential causal regulatory network for prognosis-relevant genes

To understand the relations among all of the prognosis relevant genes and to help identify the master prognosis-influencing genes, a potential causal regulatory network was estimated (see Materials and methods). Nodes in the network were all with significant prognostic impacts ($P<0.01$ and $|coef| > 0.2$ based on Cox survival analysis) (Figure 3A). The causal structure of this network was inferred from the transcriptional expressions of genes across TCGA-LUAD patients, where directed edges describe the identified direct causal effects, and bi-directed edges represent uncertainty in the constructed network (23). According to this network, we found that most of the nodes with the same type of prognosis impacts (either favourable or un-favourable) gathered together while nodes with reversed impacts were relatively separated in the network. Since the network described the potential causal structure, the survival impacts of many nodes may be in-directly generated from their up-stream or down-stream items. Consequently, it is more important to identify which prognosis-relevant genes play the master roles rather than just examining the prognosis relevance.

To distinguish the master regulators, we further investigated on the node importance in the causal network. Here we calculated the summarized node degree by subtracting the number of in-coming edges from the number of out-coming edges for certain node. We found that the summarized degrees for a large portion of nodes were within the range of -5 to 5 (Figure 3B). The other nodes which were with larger absolute values of summarized degrees were hub nodes of the network (only names of the hub nodes were displayed on Figure 3A). These nodes are more likely to play master roles in influencing LUAD prognosis, since the expression alterations of multiple prognosis relevant genes were highly associated with these hub nodes. For instance, GAPDH, CCNA2, WWP2 and PSMD2 were all hub nodes (summarized degree >5 or summarized degree <-5), and they may lead to remarkable prognosis influences by either passing its alteration to a series of down-stream partners (see sub-graphs for GAPDH and CCNA2, Figure 3C) or gathering the influences from a collection of up-stream elements (see sub-graphs for WWP2 and PSMD2, Figure 3C). The network structure may also indicate potential regulatory relationships between nodes. GAPDH, ALDOA and PKM2 all play functions in glycolysis (37), here, we also observed that they aggregated in the same sub-graph, implying that the data-driven network may suggest meaningful biological associations.

Clustering analysis reveals LUAD-subtypes

To understand the whole molecular and prognostic impact generated by the above identified master prognosis impacting genes, we clustered the LUAD patients into two groups (termed as C1 and C2 respectively) based on the mRNA-level expressions of all of the hub nodes in the causal regulatory network. We further evaluated the importance of each master gene, the top-10 important master genes for the two subtypes included *CCNA2*, *CBX7*, *TMEM48*, *SPC25*, *GAPDH*, *WDHD1*, *PSMD2*, *ERO1L*, *DDX52*, *ARNTL2* (Figure 4A, only the top-50 important genes were shown in the central heat map). These two clusters showed significantly different expressional patterns, especially for genes in the mTOR signaling pathway, lysosome, and PPRA signaling pathways (Figure 4B).

Meanwhile, potential SMGs which may be related with the expressional differences between the two identified clusters were also identified. Mutations of *SMARCA4*, *KEAP1*, *TP53* or *COL11A1* were significantly enriched in C1 (Figure 4A). These mutations were related with the significantly differential expressions for specific genes across the two clusters. For instance, *GAPDH* was highly expressed in C1, and its expression levels in patients with mutations in *TP53*, *KEAP1* or *SMARCA4* were significantly higher than wild type patients (Figure 4A). Notably, we also observed that patients in C1 were with significantly worse prognosis than C2 (Figure 5A, P-value = $2.9*10^{-8}$), and the significance was more remarkable than those obtained from random gene sets with the same number of genes (Figure 5B). Taken together, the master prognosis regulating genes help identify two meaningful subtypes for LUAD which showed significantly differential patterns in genomic and transcriptional levels.

Validating the prognosis differences between the identified LUAD subtypes

To estimate the reliability and robustness of the identified LUAD subtypes, we utilized additional two GEO datasets to further validate the expressional and prognostic patterns of these two subtypes. Firstly, a subtype predictor was trained based on the expressional profiles of the identified top-10 important master genes in the TCGA-LUAD cohorts (see details in Materials and methods). Based on this predictor, other patients having similar expressional profiles with C1 or C2 will be annotated with the corresponding sub-type labels. Then, this predictor was applied on two independent LUAD cohorts (GSE30219 and GSE31210). Thus, patients in the independent cohort were also annotated into the two subtypes (C1 and C2) based on their expressional profiles, and survival analysis showed these two groups were also with significantly differential survival outcomes just as observed in TCGA-LUAD, where C1 was with significantly worse prognosis than C2 (Figure 5C and 5D). This verifies the robustness of the identified expression-based LUAD subtypes in independent cohorts.

Promising Drugs for the identified LUAD-subtypes

Since the survival differences were highly related with the expressional alterations in the identified clusters, drugs targeting on the most important genes for classifying the two subtypes may generate distinctive effects on the two subtypes. Based on this hypothesis, we attempted to obtain potential drug-gene interactions for the top-30 important genes (see Materials and methods). As results, we found a collection of drugs like Estriol, Ethinyl and Folic acid (28) may be associated with genes like *GAPDH*, *CCNA2* and *PSMD2* (Figure 5E, and three of them were among the top-10 important genes for the two subtypes) which may be highly contributable to the survival differences between the two identified subtypes. For LUAD patients, more attention should be paid on these drugs since patients from different subtypes may have distinctive responses to these drugs.

Discussion

LUAD is one of the most common cancer types, threatening the human health around the world. The development of targeted therapies, especially those targeting on EGFR (3) and ALK (5), have promoted the treatment of some LUAD patients, however, the highly heterogeneity of LUAD makes the benefits of these therapies limited to few patients. To obtain a better understanding of the molecular patterns underlying LUAD, we put-forward a multi-omics based investigation on LUAD. Through this study, prognostic meaningful lung adenocarcinoma subtypes which are independent of EGFR and ALK mutations and the relevant mutational and expressional profiles were identified.

With the development of high-throughput biological and chemical technology, a great deal of omics-data is accumulated to help describe the molecular mechanisms of different types of cancers. Owing to omics-data measured for LUAD cohorts (15, 34, 38), a large number of significantly mutated, prognosis-relevant or differentially expressed genes for LUAD can be identified. However, the latent relationships between these interesting items were rarely revealed. In this study, we not only identified the prognosis-relevant genes, but also constructed the potential causal regulating structures among these genes, thus identifying which genes are more likely to play master roles in influencing the LUAD patients' prognosis in the transcriptional level. Based on these master genes, we also identified two potential LUAD subtypes. The poor survival rate of one subtype may be related with mutations in SMARCA4, KEAP1, TP53 and COL11A1. Low expression of SMARCA4 has been reported to be significantly associated with poor prognosis and can be served as a predictive biomarker of increased sensitivity to platinum-based therapies (39). Here, the significant mutations of SMARCA4 were also related with the poor survival rate of one LUAD subtype (Figure 4), and the mutations may lead to decreased expressions of SMARCA4 (Figure S1). Similarly, KEAP1 (40), TP53 (41) and COL11A1 (33) have all been

reported to play roles in LUAD. Co-occurrence of these SMGs in the poor survival subtype implies that the differential prognosis between the two subtypes is not simply the result of one specific gene but a collection of meaningful genes. Meanwhile, the molecular mechanism underlying the two sub-types is associated with multiple down-stream pathways, e.g., mTOR signaling pathway and lysosome.

An important issue of omics-based cancer studies is whether the revealed results can be re-discovered in the other independent cohorts despite cancer heterogeneity or sample biases. Here, based on the expressional profile of master genes, the two identified subtypes were consistent in multiple independent cohorts, confirming the robustness of the identified subtypes which showed significant differences both molecularly and clinically. The robustness of the subtypes also imply that the causal regulatory network based method help identify the most influential genes. These results can provide an alternative way to classify LUAD patients and supply valuable references on selecting the most beneficial treatments for specific type of LUAD.

A limitation of this study is that most of the calculated relationships were significant in the statistical level. It is unavoidable that false positives are mixed into these statistical relations, e.g., the causal regulating effects. However, these findings still provide remarkable data resources, which may promote the discovery of promising molecular mechanisms underlying LUAD in a less time-and resource-consuming way. In the future research, more efforts will be put into validating these potential relations.

Declarations

Acknowledgement

We thank the ZhenHe Co for helpful discussion and critical comments on the manuscript. This study is supported by The Second Affiliated Hospital of Dalian Medical University internal funds.

Author Contributions

Z.L. and T.L. designed and performed most of the data analysis and wrote the manuscript.

Conflict of interest

The authors declared no conflict of interest.

References

1. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. Adv Exp Med Biol. 2016;893:1-19.
2. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. Eur Respir J. 2016;48(3):889-902.
3. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304(5676):1497-500.
4. Muller IB, de Langen AJ, Giovannetti E, Peters GJ. Anaplastic lymphoma kinase inhibition in metastatic non-small cell lung cancer: clinical impact of alectinib. Onco Targets Ther. 2017;10:4535-41.
5. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363(18):1693-703.
6. Drilon A, Wang L, Hasanovic A, Suehara Y, Lipson D, Stephens P, et al. Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas. Cancer Discov. 2013;3(6):630-5.
7. Bergethon K, Shaw AT, Ou SH, Katayama R, Lovly CM, McDonald NT, et al. ROS1 rearrangements define a unique molecular class of lung cancers. J Clin Oncol. 2012;30(8):863-70.
8. Riely GJ, Brahmer JR, Planchard D. A randomized discontinuation phase II trial of ridaforolimus in non-small cell lung cancer (NSCLC) patients with KRAS mutations. J Clin Oncol. 2012;30(15S):abstract 7531.
9. Ou SH, Kwak EL, Siwak-Tapp C, Dy J, Bergethon K, Clark JW, et al. Activity of crizotinib (PF02341066), a dual mesenchymal-epithelial transition (MET) and anaplastic lymphoma kinase (ALK) inhibitor, in a non-small cell lung cancer patient with de novo MET amplification. J Thorac Oncol. 2011;6(5):942-6.
10. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543-50.
11. 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.
12. Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013;5(186):186ra66.
13. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med. 2007;356(1):11-20.

14. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, et al. Three-gene prognostic classifier for early-stage non small-cell lung cancer. J Clin Oncol. 2007;25(35):5562-9.
15. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-25.
16. Floyd SR, Pacold ME, Huang Q, Clarke SM, Lam FC, Cannell IG, et al. The bromodomain protein Brd4 insulates chromatin from DNA damage signalling. Nature. 2013;498(7453):246-50.
17. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
18. 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(7):e47.
19. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.
20. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.
21. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
22. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Softw. 2012;46(11):1-17.
23. Kalisch M, Machler M, Colombo D, Maathuis MH, Buhlmann P. Causal Inference Using Graphical Models with the R Package pcalg. J Stat Softw. 2012;47(11):1-26.
24. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572-3.
25. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):5.
26. Sean D, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846-7.
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw. 2011;39(5):1-13.
28. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. Nat Methods. 2013;10(12):1209-+.
29. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat

Genet. 2013;45(10):1113-20.

30. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
31. Chong IW, Chang MY, Chang HC, Yu YP, Sheu CC, Tsai JR, et al. Great potential of a panel of multiple hMTH1, SPD, ITGA11 and COL11A1 markers for diagnosis of patients with non-small cell lung cancer. Oncol Rep. 2006;16(5):981-8.
32. Rizou T, Perlikos F, Lagiou M, Karaglani M, Nikolopoulos S, Toumpoulis I, et al. Development of novel real-time RT-qPCR methodologies for quantification of the COL11A1 mRNA general and C transcripts and evaluation in non-small cell lung cancer specimens. J BUON. 2018;23(6):1699-710.
33. Shen L, Yang M, Lin Q, Zhang Z, Zhu B, Miao C. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. Oncol Rep. 2016;36(2):877-85.
34. van Buul JD, Allingham MJ, Samson T, Meller J, Boulter E, Garcia-Mata R, et al. RhoG regulates endothelial apical cup assembly downstream from ICAM1 engagement and is involved in leukocyte trans-endothelial migration. J Cell Biol. 2007;178(7):1279-93.
35. Zhang X, Berger FG, Yang J, Lu X. USP4 inhibits p53 through deubiquitinating and stabilizing ARF-BP1. Embo j. 2011;30(11):2177-89.
36. Guo W, Ma J, Pei T, Zhao T, Guo S, Yi X, et al. Up-regulated deubiquitinase USP4 plays an oncogenic role in melanoma. J Cell Mol Med. 2018;22(5):2944-54.
37. Gill KS, Fernandes P, O'Donovan TR, McKenna SL, Doddakula KK, Power DG, et al. Glycolysis inhibition as a cancer treatment and its role in an anti-tumour immune response. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2016;1866(1):87-105.
38. Lin P, Ye RD. The lysophospholipid receptor G2A activates a specific combination of G proteins and promotes apoptosis. J Biol Chem. 2003;278(16):14379-86.
39. Felip E, Gridelli C, Baas P, Rosell R, Stahel R, Panel M. Metastatic non-small-cell lung cancer: consensus on pathology and molecular tests, first-line, second-line, and third-line therapy: 1st ESMO Consensus Conference in Lung Cancer; Lugano 2010. Ann Oncol. 2011;22(7):1507-19.
40. Choi M, Kadara H, Zhang J, Parra ER, Rodriguez-Canales J, Gaffney SG, et al. Mutation profiles in early-stage lung squamous cell carcinoma with clinical follow-up and correlation with markers of immune function. Ann Oncol. 2017;28(1):83-9.
41. Nelson HH, Wilkojmen M, Marsit CJ, Kelsey KT. TP53 mutation, allelism and survival in non-small cell lung cancer. Carcinogenesis. 2005;26(10):1770-3.

Figures

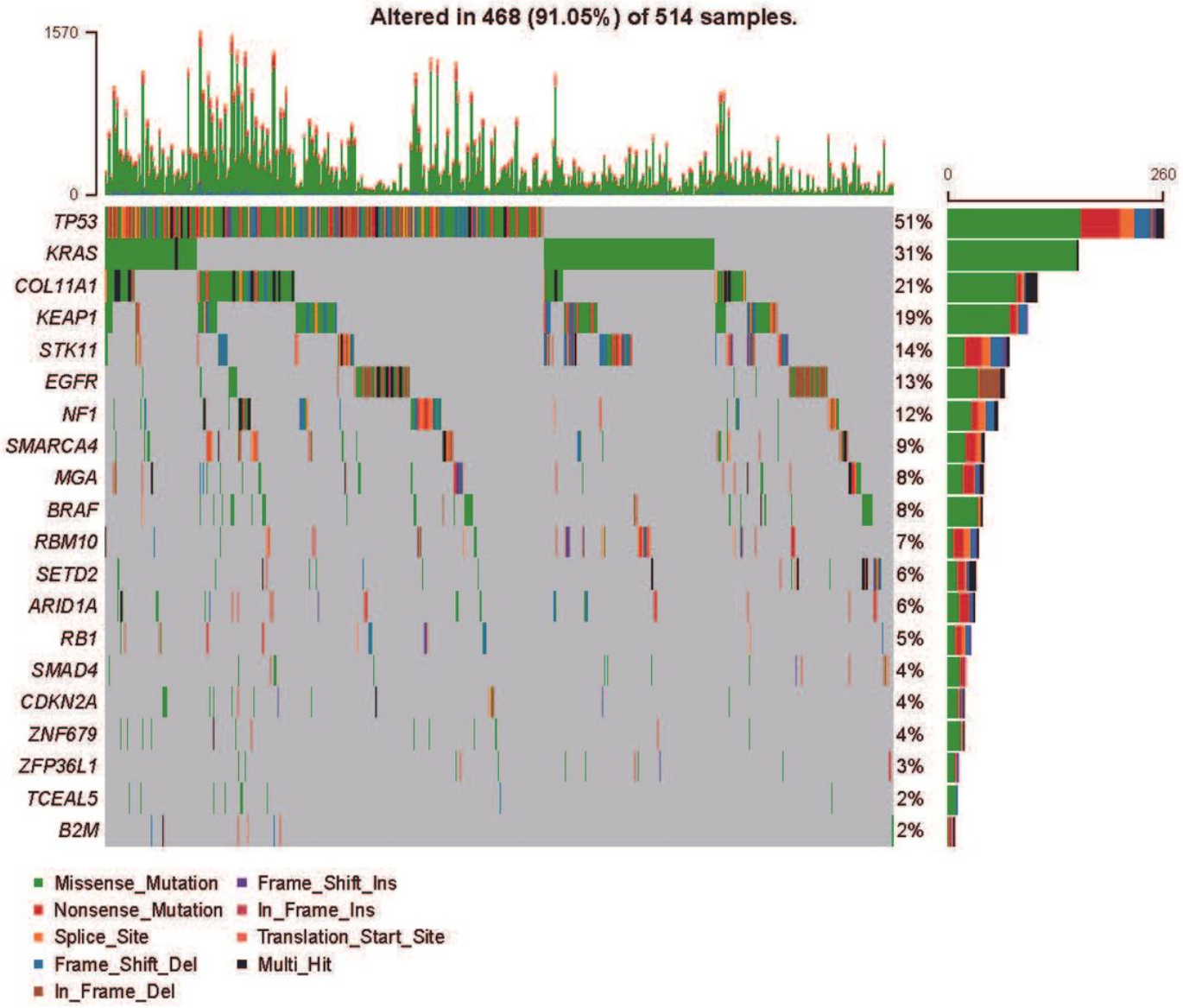

Figure 1

Significant somatic gene mutations in lung cancer. Different colors stand for different mutation types.

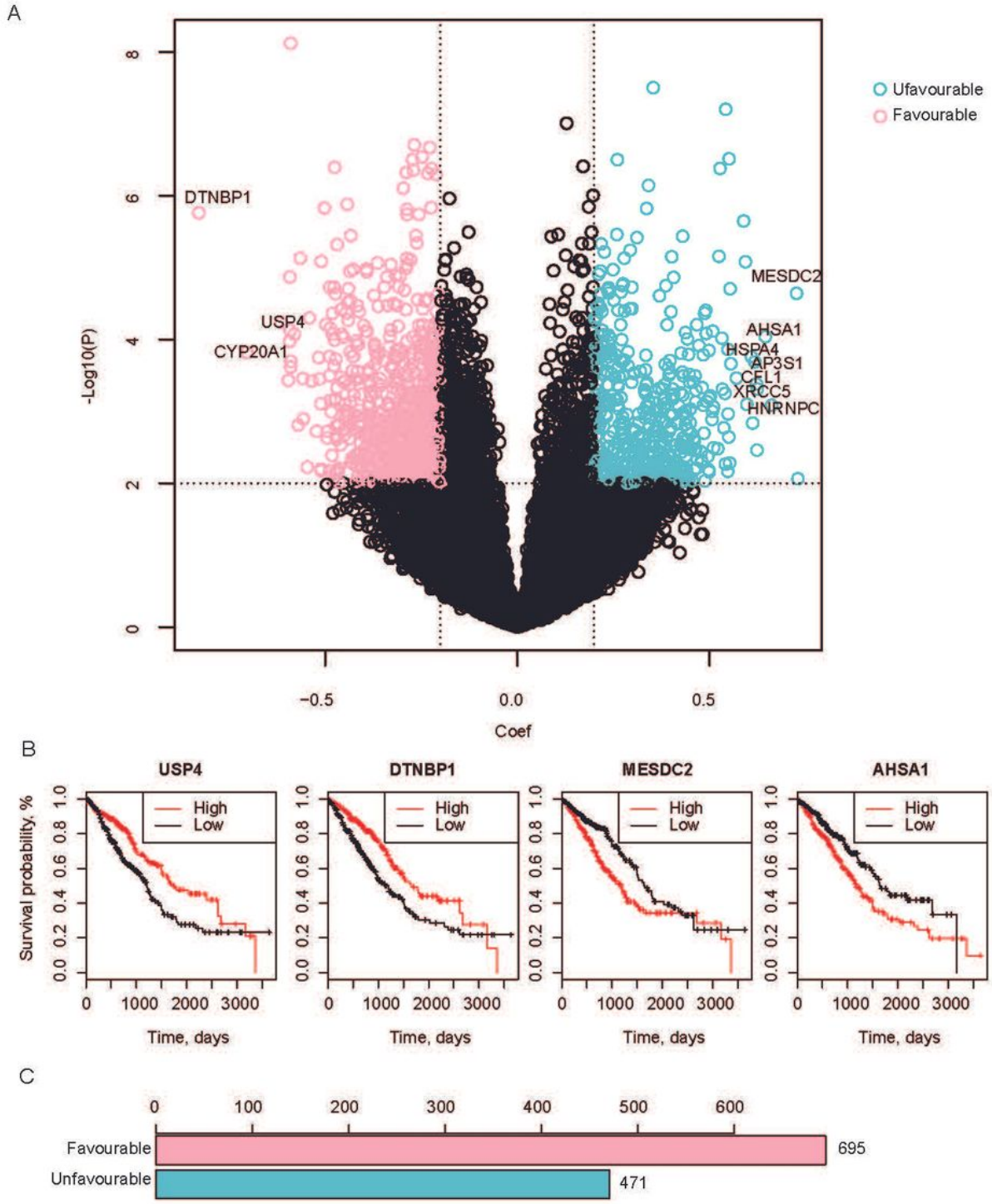

Figure 2

Genome wide prognostic analysis results. (A) A volcano plot showing the genome wide prognosis analysis results. The vertical and horizontal axis respectively represent the –log10 (p) and regression coefficient value (coef) got from the Coxph survival analysis. Each data point represents the Coxph analysis result of one gene in the mRNA level. coef < 0 means hazard ratio (HR) < 1 and coef >0 means HR > 1. Strictly, only genes with p<0.01 and coef >0.2 were regarded as unfavourable for prognosis while

genes with p<0.01 and coef <-0.2 were regared as favourable. (B) The number of significant favourable and unfavourable genes. (C) Survival curves for examples of favourable and unfavourable genes.

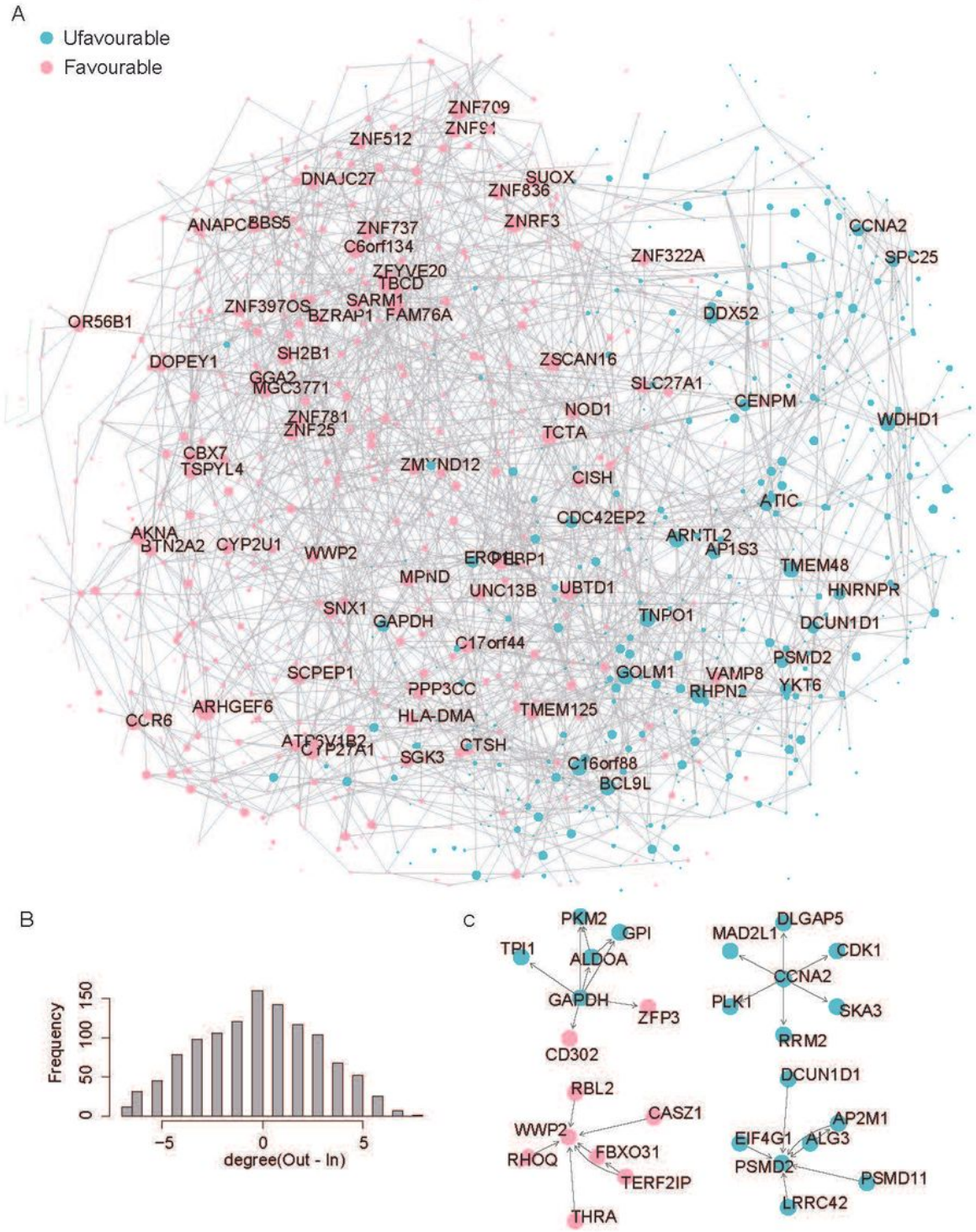

Figure 3

A causal regulatory network for prognosis relevant network. (A) The prognosis relevant causal regulatory network, where nodes represent genes and edges represent potential causal correlations. For clarity, only

nodes with degrees larger than 5 are displayed. (B) Degree distribution of the nodes. (C) Representative sub-graphs.

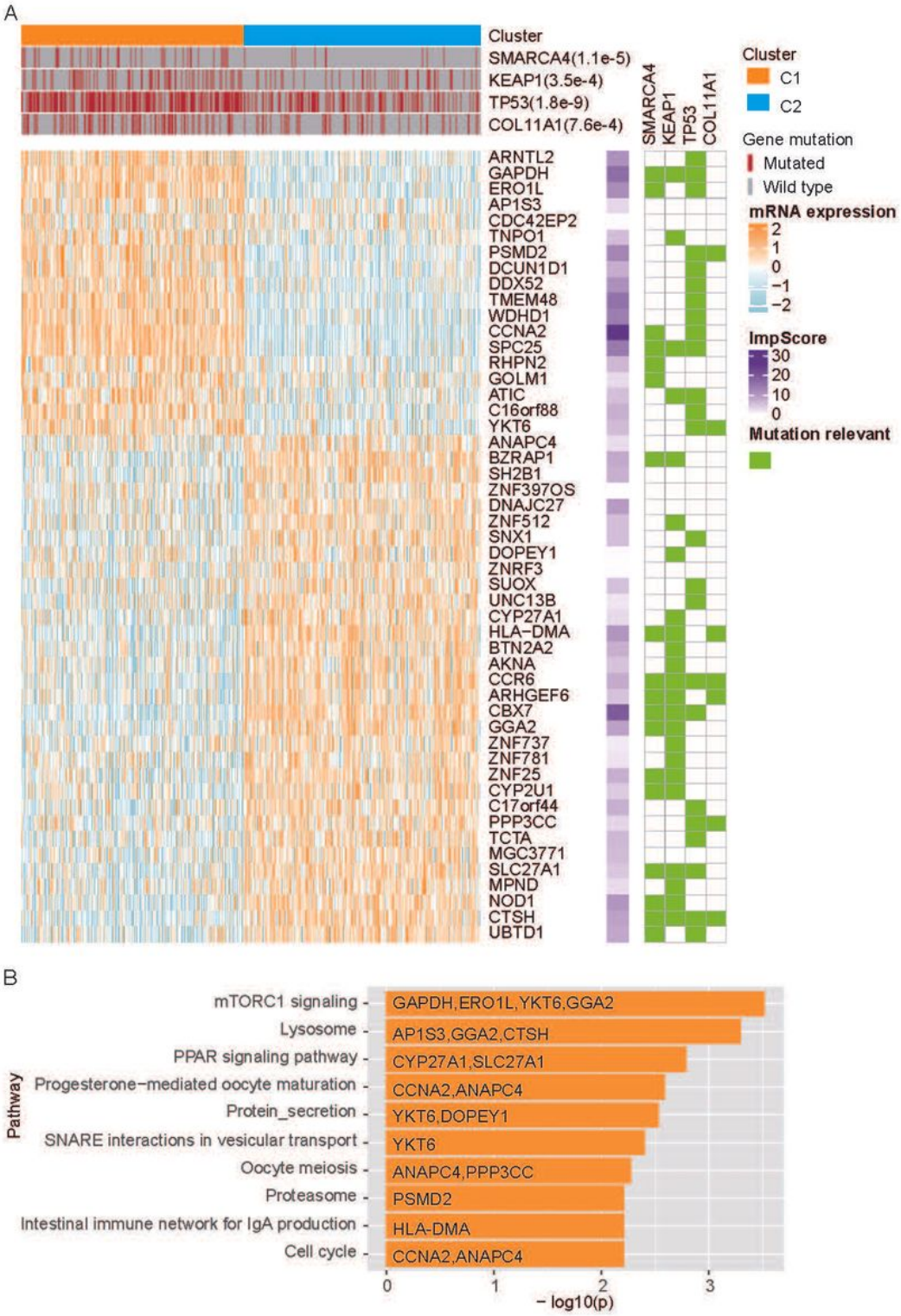

Figure 4

Clustering patients into subtypes based on high degree nodes in the causal regulator network. (A) A heatmap showing the expression profiles of the top-50 important genes for the clustering results. Above the heatmap, annotations about the clustering labels and SMGs which showed significant enrichment in

one of the cluster were shown. On the right side of the heatmap, the first column represents the importance score of each gene, and the other columns stand for the relevancies between the genes and SMGs where a green grid means that expressions of the gene is with significant difference between the SMG mutated and wild type samples. (B) Pathway enrichment analysis results for the top-50 important genes.

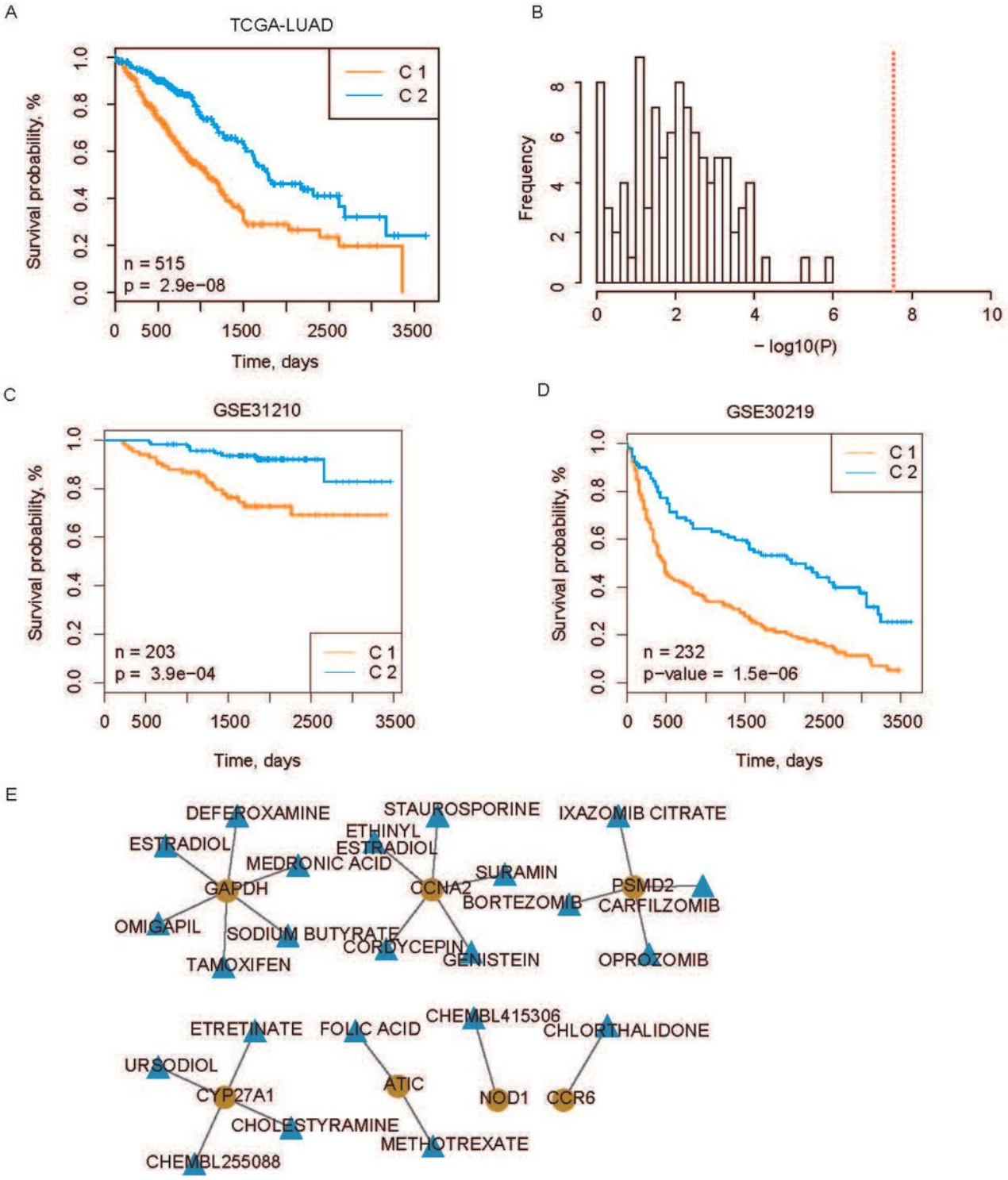

Figure 5

Validate prognosis relevant expression profiles in independent LUAD cohorts. (A) A Kaplan-Meier (KM) plot of the survival curves for the two estimated clusters. (B) Distribution of the –log10 transformed log-rank p-values for comparing the survival rates between two clusters identified based on randomly-selected genes. The red line represents the –log10 (p) got in (A). (C-D) Kaplan-Meier (KM) plots of the survival curves for two predicted patient clusters in GSE31210 (C) and GSE30219 (D). (E). Drug interaction network. Blue and yellow nodes respectively represent drugs and genes. Edges mean the gene/protein is one of the reported targets for the connected drug.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.pdf
- supplement2.docx