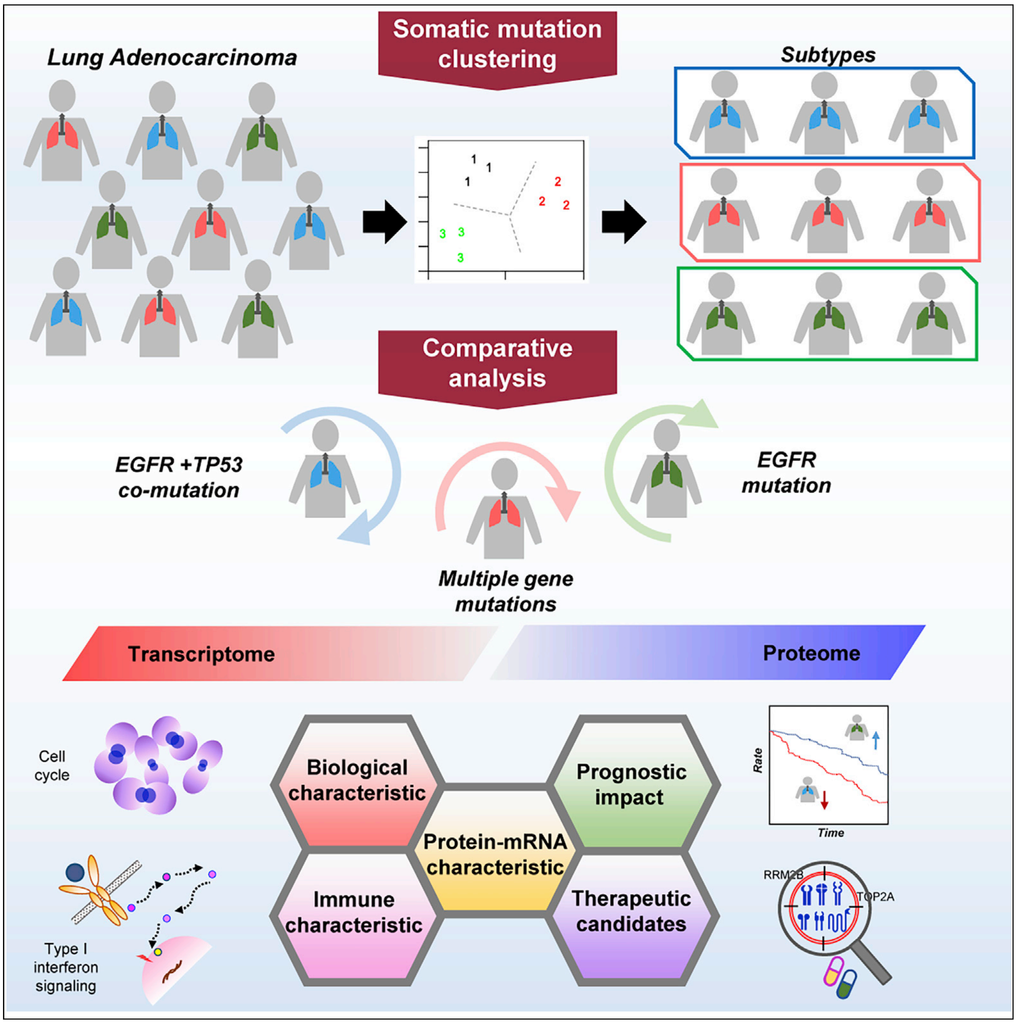
iScience

CellPress
OPEN ACCESS

Article

Somatic mutation subtypes of lung adenocarcinoma in East Asian reveal divergent biological characteristics and therapeutic vulnerabilities



Wai-Kok Choong, Ting-Yi Sung

tsung@iis.sinica.edu.tw

Highlights

Comprehensive clustering analysis reveals three somatic mutation subtypes

Prognosis of *EGFR*^{mut}/ *TP53*^{mut} subtype is worse than *EGFR*^{mut} subtype

EGFR^{mut}/*TP53*^{mut} subtype shows IFN signaling and antigen processing pathway signatures

Proteome analysis identifies druggable proteins and candidates for drug repositioning

Choong & Sung, iScience *24*, 102522
June 25, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102522

iScience

CellPress
OPEN ACCESS

Article

Somatic mutation subtypes of lung adenocarcinoma in East Asian reveal divergent biological characteristics and therapeutic vulnerabilities

Wai-Kok Choong[1] and Ting-Yi Sung[1,2,*]

SUMMARY

Lung adenocarcinoma (LUAD) patients in East Asia predominantly harbor oncogenic *EGFR* mutations. However, there remains a limited understanding of the biological characteristics and therapeutic vulnerabilities of the concurrent mutations of *EGFR* and other genes in LUAD. Here, we performed comprehensive bioinformatics analyses on 88 treatment-naïve East Asian LUAD patients. Based on somatic mutation clustering, we identified three somatic mutation subtypes: *EGFR* + *TP53* co-mutation, *EGFR* mutation, and multiple-gene mutation. A proteogenomic analysis among subtypes revealed varying degrees of dysregulation in cell-cycle-related and immune-related processes. An immune-characteristic analysis revealed higher PDL1 protein expression in the *EGFR* + *TP53* co-mutation subtype than in the *EGFR* mutation subtype, which may affect the therapeutic efficacy of anti-PD-L1 therapy. Moreover, integrating known and potential therapeutic target analysis reveals therapeutic vulnerabilities of specific subtypes and nominates candidate biomarkers for therapeutic intervention. This study provides new biological insight and therapeutic opportunities with respect to *EGFR*-mutant LUAD subtypes.

INTRODUCTION

For several years, lung cancer has been the top cause of cancer mortality in the United States and worldwide (Bray et al., 2018; Siegel et al., 2019). Among the two main types of lung cancer, non-small cell lung cancer (NSCLC) occurs much more frequently in populations in which lung adenocarcinoma (LUAD) is the most common histologic subtype (Chen et al., 2014). Commonly mutated oncogenic drivers in LUAD include *KRAS*, *EGFR*, *ALK*, and *BRAF* genes, which are reported to show high heterogeneity in LUAD (Herbst et al., 2018; Yang et al., 2020). Some of these mutated genes, such as *KRAS* and *EGFR*, are usually observed to be mutually exclusive in a majority of LUAD patients (CancerGenomeAtlasResearchNetwork, 2014; Ding et al., 2008; Jordan et al., 2017). Furthermore, the frequencies of *KRAS* and *EGFR* mutations differ between East Asian and Caucasian LUAD patients, where *EGFR* mutation is predominant in East Asians and *KRAS* in Caucasians (Gahr et al., 2013; Shigematsu et al., 2005; Wu et al., 2008). Although treatment of LUAD in clinical practice usually involves drugs to inhibit specific targeted oncogenic drivers, clinical trials show that it is rare for patients to experience complete remission (Rosell et al., 2012; Solomon et al., 2014; Yang et al., 2015; Zhou et al., 2011). Moreover, the molecular homogeneity of an oncogenic driver subgroup is usually low (Chen et al., 2017, 2020a; Lv and Lei, 2020; Zhang et al., 2019). This suggests that the single oncogenic driver model is insufficient for decoding the heterogeneity of LUAD.

In several omics-based studies, LUAD heterogeneity is investigated by adopting the molecular subtyping approach in which unsupervised clustering methods are applied to mRNA and protein expressions; the results provide novel insights into molecular heterogeneity (CancerGenomeAtlasResearchNetwork, 2014; Chen et al., 2017; Chen et al., 2020a; Gillette et al., 2020). However, the various molecular subtypings of LUAD obtained from such studies are generally incompatible, suggesting that tumor heterogeneity involves other unknown factors. For example, based on East Asian LUAD cohorts, Xu et al. (Xu et al., 2020) report three proteomic subtypes–EM-H subtype (environment and metabolism high), PP subtype (proliferation and proteasome function), and mixed subtype–and Chen et al. (Chen et al., 2020b) propose five proteomic subtypes highly associated with TNM stage classification. In contrast, studies based on genomic sequencing show that multiple non-random co-occurring mutations affect biological pathways and clinical

[1] Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan

[2] Lead contact

*Correspondence: tsung@iis.sinica.edu.tw

https://doi.org/10.1016/j.isci.2021.102522

outcomes in NSCLC (Blakely et al., 2017; CancerGenomeAtlasResearchNetwork, 2014; Ding et al., 2008; Jordan et al., 2017). For instance, the respective mutations of *STK11* and *KEAP1* genes frequently co-occur with the *KRAS* mutation in NSCLC patients (Arbor et al., 2018; Scheffler et al., 2019). Notably, a new molecular classification model that accounts for the impact of co-occurring mutations has been proposed and is potentially more efficacious than the single oncogenic driver model to decode LUAD heterogeneity (Skoulidis and Heymach, 2019). However, the impact of co-occurring mutations on omics-based expression profiles of LUAD remains relatively unexplored.

In our previous study (Chen et al., 2020b) we reported a comprehensive proteogenomic analysis of paired tumor and adjacent normal tissues acquired from Taiwan LUAD patients and revealed molecular characterization of pathogenesis and progression in an early stage non-smoking LUAD cohort. However, co-occurring mutation patterns that may exist in this East Asian LUAD cohort still need in-depth examination for a better understanding of their impact on biological functions and therapeutic vulnerabilities. Therefore, we are motivated to conduct comprehensive bioinformatics analyses for thoroughly inspecting patterns of co-occurring mutations in our East Asian LUAD cohort as well as the impact thereof.

In this paper, we present the first somatic mutation subtyping analysis of an East Asian cohort based on possible co-mutation patterns of 88 patients. Despite the sparseness and heterogeneity of somatic mutation profiles, this cohort can be classified into three distinct mutation pattern subtypes. Transcriptomic and proteomic analyses show that biological processes such as cell cycles of these subtypes are significantly discordant. To discover the hallmark proteins of each subtype suitable for clinical applications, we further examine existing drug targets and provide potential testable targets for therapy specific to the somatic mutation subtypes. Our findings highlight the importance of identifying LUAD somatic mutation subtypes, which reveal diversified molecular and clinical characterization.

RESULTS

Clinical, mutation features and multi-omics profiling of Taiwan LUAD cohort

In the multi-omic, somatic mutation, and clinical data of 88 treatment-naïve LUAD patients obtained from our previous study (Chen et al., 2020b), the transcriptomic and proteomic expression profiles contain 30,155 genes and 13,458 proteins, respectively. In this cohort, 44.32%, 35.23%, and 20.45% were diagnosed as stage IA, stage IB, and stage II–IV, respectively. The average age of all patients was 63.47, and a majority (86.36%) of the patients were never-smokers (Figure 1A; Table S1A).

Among the 88 LUAD paired tumor-adjacent normal tissue samples, a total of 10,054 genes contained at least one somatic mutation variant event derived from whole-exome sequencing (WES) analysis. Overall, only four genes with somatic mutations were observed in tumors of more than 20% of the samples, namely, *EGFR* (84.09%), *TP53* (29.55%), *TTN* (25.0%), and *MUC4* (20.45%) (Figure 1A; Table S1A), where *EGFR* mutation occurred dominantly in our LUAD patients. Among the top10 frequent somatic mutation genes in our cohort, the frequencies of the eight genes–*EGFR*, *TP53*, *TTN*, *MUC4*, *MUC16*, *FLNC*, MUC12, and *RBM10*–were significantly different (Fisher's exact test, Benjamini-Hochberg [BH] adjusted [adj.] $p < 0.05$) from the TCGA LUAD cohort (n = 574) (BITGDACenter, 2016) (Figure S1A; Table S1B). The results show that the somatic mutation feature of our LUAD cohort is different from that of the TCGA cohort.

Somatic mutation subtyping of East Asian LUAD cohort

Genomic alteration of each gene is generally an independent event involved in interference with a biological system. However, concurrent genomic alterations of multiple genes may cause additional interferences, as compared with a single gene. Previous studies (Arbor et al., 2018; Canale et al., 2017; Labbe et al., 2017; Nahar et al., 2018; Skoulidis et al., 2015) show that concurrent somatic mutations of multiple genes affect the clinical outcome of diseases. To identify possible concurrent somatic mutation patterns, we employed consensus clustering with partitioning around medoids to categorize our cohort based on the somatic mutation gene matrix. To optimize efficiency, for clustering we adopted those 20 genes with the top mutation frequencies, instead of all of the mutation genes (STAR Methods, Figures S1B–S1E and S2) and identified three subtypes: *G1* (23 patients) with *EGFR* + *TP53* mutation, *G2* (58 patients) mainly with *EGFR* mutation alone, and *G3* (7 patients) with multiple-gene mutations (Figure 1B; Table S1C). All three subtypes–*EGFR* + *TP53* mutation, *EGFR* mutation, and multiple-gene mutation—were shown to be statistically insignificant with gender, age, smoking, and stage features, respectively (right-tailed Fisher's exact test, Bonferroni adj. $p > 0.05$) (Table S1D). Observing that G2 contained 11 (19%) tumor

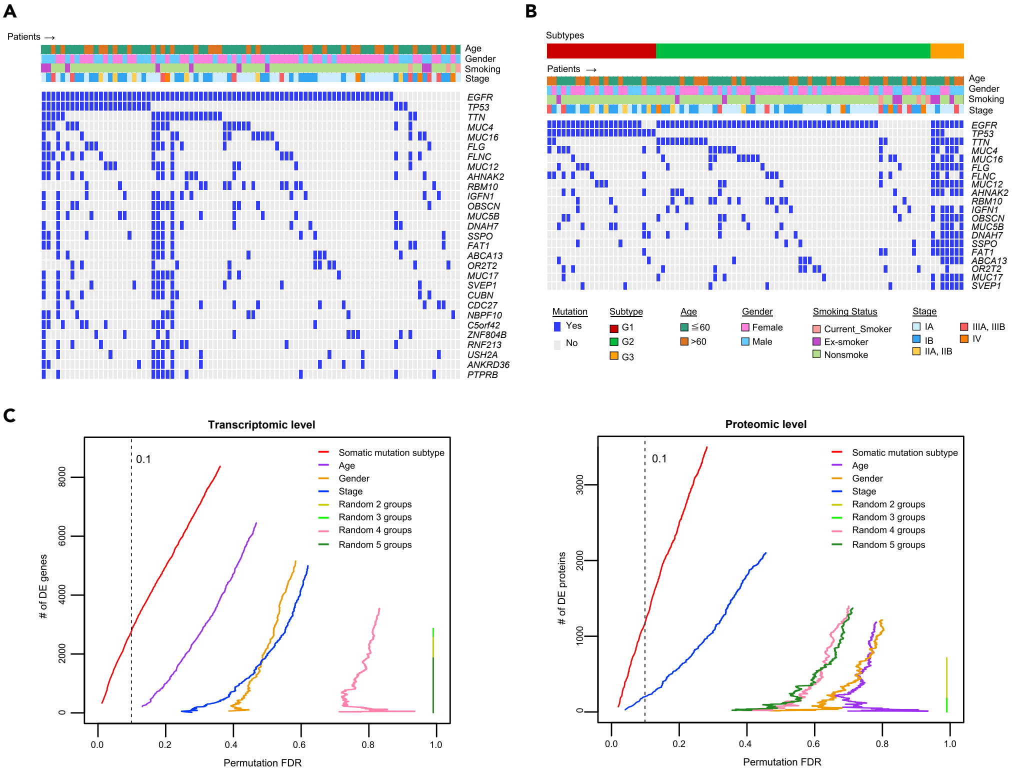


Figure 1. Somatic mutation profile and subtyping of 88 Taiwan LUAD cohort

(A) Somatic mutation profile. Clinical features and 29 somatic mutation genes (with more than 12% occurrence frequency in 88 patients) annotated for each patient.

(B) Three somatic mutation subtypes obtained by using consensus clustering on top20 frequent somatic mutation genes.

(C) Comprehensive comparison of mRNA (left) and protein (right) expression profiling in different clusterings of the cohort based on somatic mutation subtyping, three clinical features, and four random sampling, respectively. The curves derived from differential expression analyses show the number of genes/proteins with ANOVA $p < 0.05$ at different levels of permutation FDR.

See also Figures S1 and S2 and Table S1.

samples without *EGFR* mutation, we examined the results of clustering into k = 4, 5 groups, in which these 11 samples were still clustered together with most of the other 47 samples with *EGFR* mutation in one group and not mixed with other groups. These clustering results strongly suggest that the tumor samples of G2 share concurrent mutations of genes other than *EGFR*, contributing to similarity among these samples. That is, other mutation genes of these patients may contribute more similarity to the second subtype (STAR Methods).

To justify the suitability of the somatic mutation subtypes obtained from our LUAD samples for further biological discovery, we conducted a robust comparison analysis on both transcriptomic and proteomic expression profiling in terms of log2 T/N (T: abundance in tumor, N: abundance in adjacent normal tissue). First, we constructed four datasets as negative controls that randomly clustered our cohort of 88 patients into two to five groups, respectively. Moreover, we constructed three reference datasets, clustering by age (2 groups), gender (2 groups), and stage (4 groups). Then, we performed a differential expression (DE) analysis among subtypes or groups on all of the above datasets at both transcriptomic and proteomic levels. Because a rigorous DE analysis on a negative control dataset usually shows no differentially expressed genes/proteins among its randomly sampled groups, to justify the applicability of our somatic mutation subtypes for further biological discovery, we can use the curve of the number of DE genes/proteins versus various false discovery rate (FDR) level derived from the negative control and the reference datasets. Notably, based on both transcriptomic and proteomic expression data, the curves of all negative-control and most of the reference datasets are located outside the FDR <0.1 area (Figure 1C; Table S1E), except for the stage reference dataset on proteomic data (Discussion section). This means that no significant DE

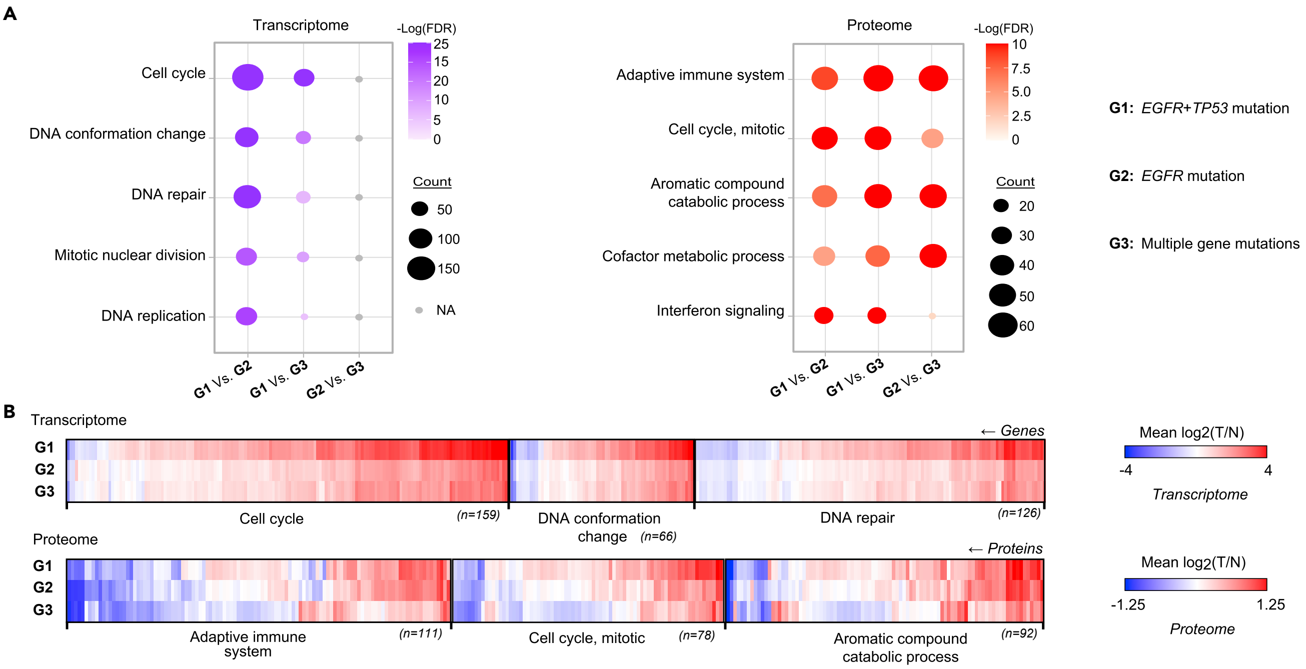


Figure 2. Commonly enriched biological processes and molecular expression

(A) Top five significantly enriched biological processes (FDR<0.05) commonly appearing in pairwise comparisons of three somatic mutation subtypes based on differentially expressed genes and proteins at transcriptome (left panel) and proteome (right panel) levels, respectively.

(B) Heatmap of differentially expressed genes (top panel) and proteins (bottom panel) in top three enriched biological processes.

See also Figures S3–S5 and Table S2.

genes/proteins among groups in these datasets pass an FDR of 0.1, thus revealing the high homogeneity of these groups at mRNA and protein levels. In contrast, the curves of our somatic mutation subtypes on transcriptomic and proteomic data are clearly located inside the area of FDR <0.1, which implies that the homogeneity among the three somatic mutation subtypes is significantly different at bio-molecular levels and that these DE genes/proteins may be involved in biological activities related to specific subtypes (Figure 1C). Overall, our somatic mutation subtypes are reasonable clustering results that have biological implications from multi-molecular profiling that merit further investigation.

In DE analysis, we identified 2,830 DE genes and 1,196 DE proteins (ANONA, $p < 0.015$, permutation FDR <0.1) among the three somatic mutation subtypes based on transcriptomic and proteomic profiling, respectively. Further examining the transcriptomic profiling, we obtained 2,790 DE genes between *EGFR* + *TP53* mutation and *EGFR* mutation subtypes, 718 between *EGFR* + *TP53* mutation and multiple gene mutation subtypes, and 68 between *EGFR* mutation and multiple gene mutation subtypes. Similarly, based on the proteomic profiling, we obtained 612 DE proteins between *EGFR* + *TP53* mutation and *EGFR* mutation subtypes, 640 between *EGFR* + *TP53* mutation and multiple gene mutation subtypes, and 669 between *EGFR* mutation and multiple gene mutation subtypes (Table S1F).

Biological characteristics of three LUAD somatic mutation subtypes

To explore the biological characteristics of the somatic mutation subtypes, we performed a biological process enrichment analysis using the DE genes/proteins between any two subtypes of the three subtypes (STAR Methods). At the transcriptomic level, the top five enriched processes that commonly appeared in all paired comparisons among the three subtypes were cell cycle, DNA conformation change, DNA repair, mitotic nuclear division, and DNA replication (Figure 2A; Table S2A). These enriched processes were much more enriched in G1 versus G2 comparison than in G1 versus G3. However, these processes were not significantly different between G2 and G3. Furthermore, we demonstrated the authenticity of the aforementioned enrichment results by examining the expression of the genes in the top three processes as shown in Figure 2B and Table S2B. Thus, the cell-cycle-related processes are distinct between the *EGFR* + *TP53* mutation (G1) subtype and the *EGFR* mutation (G2) subtype at the transcriptomic level.

In contrast, at the proteomic level, the top five enriched processes that commonly appeared in all paired comparisons were adaptive immune system, cell cycle (mitotic), aromatic compound catabolic process, cofactor metabolic process, and interferon signaling, which are different from those obtained from transcriptomic analysis. Notably, these five enriched processes are all significantly different in G1 versus G2, G1 versus G3, and G2 versus G3 pairwise comparisons (Figure 2A; Table S2A). As before, we demonstrated the authenticity of the above enrichment results by examining the protein expressions of the top three processes as shown in Figure 2B and Table S2B. The protein expression profiling of the immune-related, metabolism-related, and cell cycle processes is distinct between any pair of the three subtypes. Overall speaking, DE genes and proteins, respectively, enriched different biological processes in the three subtypes and also showed that more diversified biological process features are exhibited at the proteomic level than at the transcriptomic level.

At the proteomic level, we also found that five proteasome subunit proteins—PSMB1, PSMB2, PSMB3, PSME1, and PSMA6—are differentially expressed among the three subtypes, which are also involved in the adaptive immune system, cell cycle (mitotic), and aromatic compound catabolic process (Figure S3A; Table S2C). In addition, other seven proteasome subunit proteins—PSMA1, PSMA2, PSMA4, PSMA5, PSMA7, PSMB9, and PSMB10—are differentially expressed between G1 and G2 (Figure S3B; Table S2D). At the transcriptomic level, we observed 14 proteasome subunit genes—PSMA5, PSMA7, PSMB3, PSMB5, PSMB8, PSMB9, PSMC3, PSMC4, PSMD1, PSMD13, PSMD2, PSMD3, PSMD4, and PSME2—are differentially expressed between G1 and G2 (Figure S4; Table S2E). This shows that the proteasome activity level of the three somatic mutation subtypes is different: *EGFR* + *TP53* mutation > *EGFR* mutation.

In addition, we discovered six subunits—MCM2, MCM3, MCM4, MCM5, MCM6, and MCM7—of minichromosome maintenance protein (MCM) complex are significantly differentially expressed between G1 and G2 at both transcriptomic and proteomic levels (Figure S5). Moreover, the overexpression of MCM complex's subunits in various cancer types has been reported in previous studies (Das et al., 2013; Hua et al., 2014; Lau et al., 2010; Liu et al., 2018). The average log2 T/N of MCM complex expressions of the three subtypes in both omics profiles are greater than zero, implying overexpression in tumor. Furthermore, their average expression compared with adjacent normal tissue of the *EGFR* + *TP53* mutation subtype (G1) is at least 1.28-fold higher than that of the *EGFR* mutation subtype (G2) at both omics levels. In summary, the above results suggest that the cell division and DNA replication activity of the *EGFR* + *TP53* mutation subtype are higher than that of the *EGFR* mutation subtype.

Protein-mRNA expression correlation analysis

Given the abundance of the 9,009 protein-mRNA pairs in the LUAD samples (Table S3A), we investigated the correlation of mRNA and protein expressions. First, in the sample-wise protein-mRNA expression correlation analysis, we calculated the expression correlation of protein-mRNA pairs of various groups and observed that the median of the Spearman correlation coefficients of all patients and G1, G2, and G3 are 0.299, 0.346, 0.295, and 0.209, respectively (Figure 3A; Table S3B). The expression correlation between protein-mRNA pairs in G1 is significantly higher than G2 (adj. $p < 0.05$). The correlations of the all-patient group and G2 are similar mainly because G2 contains a majority of the 88 LUAD cohort. These observations demonstrate that clustering-based somatic mutation subtypes is an effective and reasonable classification of the LUAD samples. Next, we performed a gene-wise protein-mRNA expression correlation analysis for the above-mentioned four groups. The resulting medians of the Spearman correlation coefficients of all the 9,009 protein-mRNA pairs in all 88 patients and G1, G2, and G3 were 0.15, 0.19, 0.14, and 0.028, respectively (Figure 3B; Table S3C). The figure shows that the correlations of the overall protein-mRNA pairs among the three somatic mutation subtypes are significantly different. The gene-wise correlation of the overall protein-mRNA pairs in the G1 group is higher than that in G2 and G3, and G3 has the lowest correlation and exhibits a broad variation of correlations. Thus, genes corresponding to relatively high-correlated protein-mRNA pairs in each subtype may be different from other subtypes; their associated biological features may be dissimilar as well.

To discover biological pathways associated with relatively high-correlated protein-mRNA pairs, we performed gene set enrichment analysis using the KEGG pathway gene sets on the three subtypes, respectively. Using WebGestalt (Liao et al., 2019) for GSEA (Gene Set Enrichment Analysis), 18 and 21 pathways were enriched for G1 and G2, respectively (FDR <0.05) (Figure 3C; Table S3D). In contrast, for G3 no pathway was enriched passing FDR <0.05. The enriched pathways of G1 mainly belong to the immune

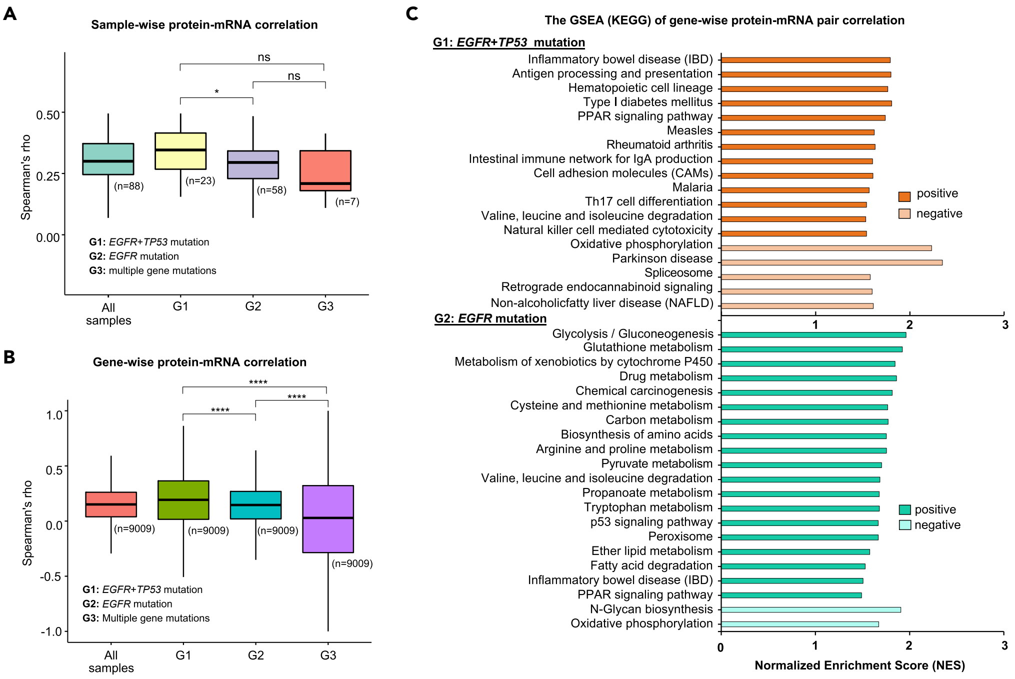


Figure 3. Protein-mRNA expression correlations across subtypes and their association with biological pathways

(A) Sample-wise protein-mRNA abundance correlation of 88 patients (p value = 0.026 by Kruskal-Wallis test) and pairwise comparison of the three subtypes (p value by Wilcoxon rank-sum test and Benjamini & Hochberg correction). Symbols indicating statistical significance are as follows: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$.

(B) Gene-wise abundance correlation of 9,909 protein-mRNA pairs in the three subtypes (p value < 0.0001) and pairwise comparison of the three subtypes.

(C) KEGG gene set enrichment analysis based on the gene-wise protein-mRNA correlation of the G1 and G2 subtypes, respectively. All KEGG pathways shown in the bar graph are statistically significantly enriched with an FDR <5%.

See also Figure S6 and Table S3.

system and human disease classes, such as antigen processing and presentation (hsa04612) and inflammatory bowel disease (hsa05321), whereas the enriched pathways of G2 are mostly associated with biomolecular metabolism processes such as glycolysis/gluconeogenesis (hsa00010) and glutathione metabolism (hsa00480). The results show that relatively high-correlated protein-mRNA pairs of the *EGFR* + *TP53* mutation and *EGFR* mutation subtypes are distinctly involved in pathways of different biological characteristics.

To further explore the unique pathway signatures derived from relatively high-correlated protein-mRNA pairs, we performed a Kolmogorov-Smirnov test (adj. $p < 0.05$) on the protein-mRNA pairs of the above enriched pathway gene sets between G1 and G2 (Table S3D). Interestingly, 11 out of 18 enriched pathways of G1, including one also enriched for G2, show significant differences between G1 and G2, and the median Spearman correlation coefficient of the genes involved in these 11 pathways of G1 is higher than that of G2 (Figure S6A; Table S3D). Although the inflammatory bowel disease pathway (hsa05321) is enriched in both G1 and G2 but shows significant difference in protein-mRNA correlation between G1 and G2, the GSEA results show that this pathway is more specific to G1 (Normalized Enrichment Score (NES): 1.79, FDR = 0.0023) than to G2 (NES: 1.5, FDR = 0.0398) (Table S3D). Therefore, these 11 pathways, including antigen processing and presentation (hsa04612) and cell adhesion molecules (hsa04514), can be regarded as the pathway signatures of the *EGFR*+*TP53* mutation subtype derived from relatively high-correlated protein-mRNA pairs. To find possible pathway signatures for G2, we observed that the 20 enriched pathways of G2 reveal similar protein-mRNA correlations, i.e., they are not significantly different (adj. $p > 0.05$) from G1 (Figure S6B; Table S3D). Hence, we can conclude that the *EGFR* mutation subtype exhibits no pathway signature.

Note that we observed that three enriched pathways in G2 are involved in the Warburg effect, and one of the three pathways is also enriched in G1. Specifically, relatively high-correlated protein-mRNA pairs are

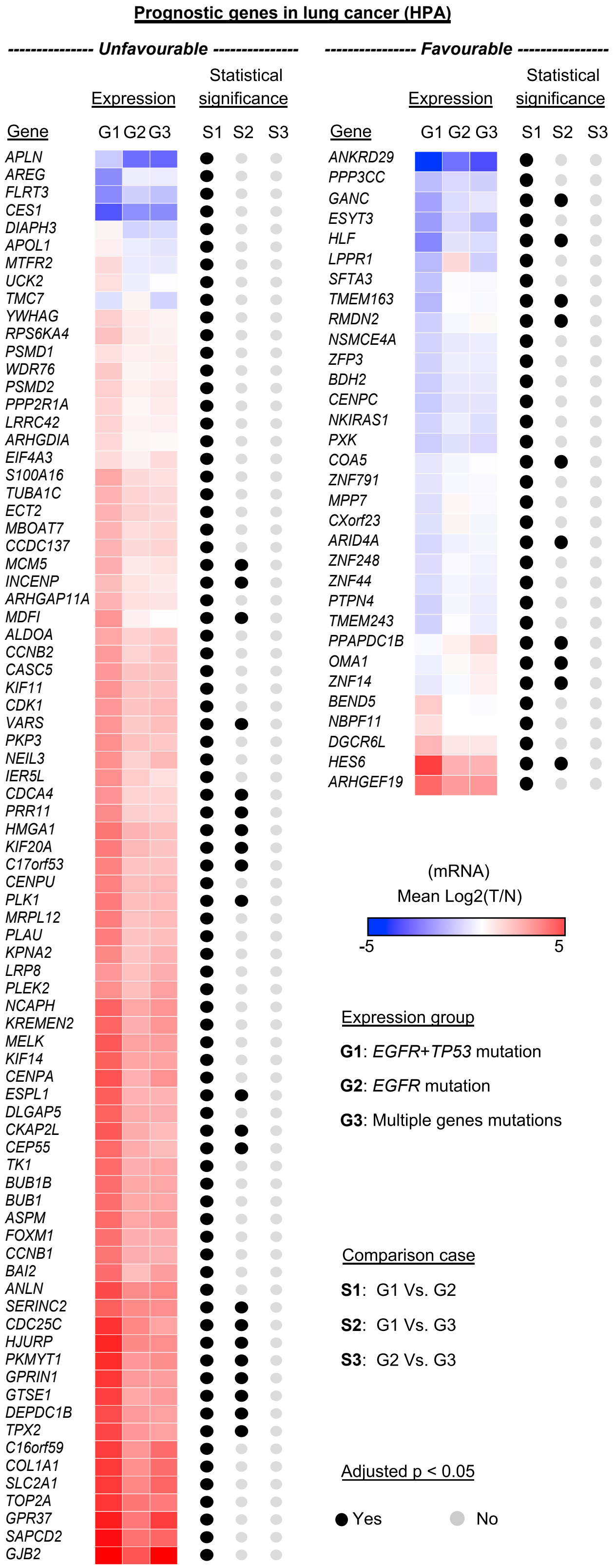

Prognostic genes in lung cancer (HPA)
--------------- Unfavourable ---------------
---------------- Favourable ----------------
Expression
Statistical significance
Gene
G1 G2 G3
S1 S2 S3
APLN
AREG
FLRT3
CES1
DIAPH3
APOL1
MTFR2
UCK2
TMC7
YWHAG
RPS6KA4
PSMD1
WDR76
PSMD2
PPP2R1A
LRRC42
ARHGDIA
EIF4A3
S100A16
TUBA1C
ECT2
MBOAT7
CCDC137
MCM5
INCENP
ARHGAP11A
MDFI
ALDOA
CCNB2
CASC5
KIF11
CDK1
VARS
PKP3
NEIL3
IER5L
CDCA4
PRR11
HMGA1
KIF20A
C17orf53
CENPU
PLK1
MRPL12
PLAU
KPNA2
LRP8
PLEK2
NCAPH
KREMEN2
MELK
KIF14
CENPA
ESPL1
DLGAP5
CKAP2L
CEP55
TK1
BUB1B
BUB1
ASPM
FOXM1
CCNB1
BAI2
ANLN
SERINC2
CDC25C
HJURP
PKMYT1
GPRIN1
GTSE1
DEPDC1B
TPX2
C16orf59
COL1A1
SLC2A1
TOP2A
GPR37
SAPCD2
GJB2
ANKRD29
PPP3CC
GANC
ESYT3
HLF
LPPR1
SFTA3
TMEM163
RMDN2
NSMCE4A
ZFP3
BDH2
CENPC
NKIRAS1
PXK
COA5
ZNF791
MPP7
CXorf23
ARID4A
ZNF248
ZNF44
PTPN4
TMEM243
PPAPDC1B
OMA1
ZNF14
BEND5
NBPF11
DGCR6L
HES6
ARHGEF19
(mRNA)
Mean Log2(T/N)
-5
5
Expression group
G1: EGFR+TP53 mutation
G2: EGFR mutation
G3: Multiple genes mutations
Comparison case
S1: G1 Vs. G2
S2: G1 Vs. G3
S3: G2 Vs. G3
Adjusted p < 0.05
● Yes
● No

Figure 4. Expression of differentially expressed prognostic genes and pairwise comparison among subtypes
Unfavorable and favorable prognostic genes are shown in the left and right panels, respectively. Listed genes are significantly differentially expressed between at least a pair of the subtypes (ANOVA, $p < 0.015$, and permutation FDR <0.1, Tukey's honest significance test, adjusted $p < 0.05$). See also Figure S7 and Table S4.

positively correlated to glycolysis/gluconeogenesis (has00010) and pyruvate metabolism (hsa00620) pathways and negatively correlated to the oxidative phosphorylation (hsa00190) pathway. This is also attested by the higher median Spearman correlation coefficients of 0.339 and 0.309 for protein-mRNA pairs involved in the glycolysis and pyruvate metabolism pathways, respectively, and a lower mean correlation of 0.012 for the oxidative phosphorylation pathway in mitochondria (Figure S6C). Because these three enriched pathways of G2 were shown above to reveal similar protein-mRNA correlations between G2 and G1, the cellular state of the *EGFR*+*TP53* mutation and *EGFR* mutation subtypes may reveal the Warburg effect, which is commonly observed in cancer cells (Bhattacharya et al., 2016; Icard et al., 2018; Lunt and Vander Heiden, 2011; Vander Heiden et al., 2009).

Estimation of prognosis of three LUAD somatic mutation subtypes

To explore the prognosis characteristics of the three LUAD somatic mutation subtypes, we first used the 2,830 DE genes to perform an enrichment analysis on the cancer-related prognostic gene categories recorded in the Human Protein Atlas (HPA) (Uhlen et al., 2010). The DE genes are significantly associated with cancer-related prognostic genes (BH adj. p = 4.02E-25) (Figure S7; Table S4A), particularly, more significantly associated with unfavorable prognostic genes (BH adj. p = 3.90E-21) than favorable prognostic genes (BH adj. p = 6.38E-04). This shows that the DE genes of the three subtypes are especially enriched in the cancer-related unfavorable prognostic gene category. Hence, DE genes that are also annotated as prognostic genes in HPA can provide clues to evaluate the prognosis of the three somatic mutation subtypes.

Among the 2,830 DE genes, there are 80 unfavorable and 32 favorable prognostic genes related to lung cancer (Figure 4; Table S4B). All of the 80 unfavorable prognostic genes are differentially expressed between G1 and G2. Moreover, 21 (26.25%) are differentially expressed between G1 and G3. Because a gene annotated as a favorable or unfavorable prognostic gene in HPA is based on the association of its high expression and clinical outcomes, we particularly investigated the expression levels of these prognostic genes between different subtypes. Notably, 75 (93.75%) of the 80 unfavorable genes are more highly expressed in G1 than in G2. Of the 32 favorable prognostic genes, all are significantly differentially expressed between G1 and G2, and 84.37% are more highly expressed in G2 than in G1. However, none of the 32 genes is significantly differentially expressed between G2 and G3. Given the above analyses of prognostic characteristics of multiple genes in lung cancer, we conclude that the *EGFR*+*TP53* mutation subtype indicates a poorer prognosis than the *EGFR* mutation and multiple gene mutation subtypes, but there is no significant prognosis difference between the *EGFR* mutation and multiple gene mutation subtypes. This observation is in line with recent clinical outcome studies on non-small cell lung cancer with concurrent *EGFR* and *TP53* mutations (Hou et al., 2019; Jiao et al., 2018; Qin et al., 2020).

PD-L1 status and immune characteristics of three somatic mutation subtypes

Cancer immunotherapy is a promising strategy for cancer treatment that modulates the host immune system to destroy tumor cells (Mellman et al., 2011). Anti-PD-1/PD-L1 immunotherapy has been approved by the US FDA and has shown significant clinical benefits for NSCLC patients (Herbst et al., 2018; Yang et al., 2020). The status of PD-L1 expression is generally used as the main predictive biomarker for anti-PD-1/PD-L1 immunotherapy response in clinical practice; high PD-L1 expression in tumor cells tends to reveal a more robust response rate (Fehrenbacher et al., 2016; Reck et al., 2016). Therefore, we further investigated the PD-L1 expression status among the three somatic mutation subtypes. PD-L1 mRNA expressions between any pair of the three subtypes show similar expression levels (Wilcoxon rank sum test, BH adj. $p \geqq 0.05$), whereas PD-L1 protein expressions are significantly different between G1 and G2 (Wilcoxon Rank Sum, Bonferroni adj. $p \leqq 0.05$) (Figure 5A, top panel). Although the three subtypes show low expression of mRNAs and proteins in tumor compared with adjacent normal tissue as evidenced by log2 (T/N) < 0, the overall PD-L1 protein expression in terms of log2 (T/N) in G1 is significantly higher than that of G2. To verify the observation of PD-L1 protein from our cohort, we also analyzed the protein expressions of PD-L1 in a large-scale LUAD study reported by CPTAC (Gillette et al., 2020) and observed that G1 is consistently and significantly higher than G2 (Wilcoxon rank sum test, $p \leqq 0.05$) (Figure 5A, bottom panel;

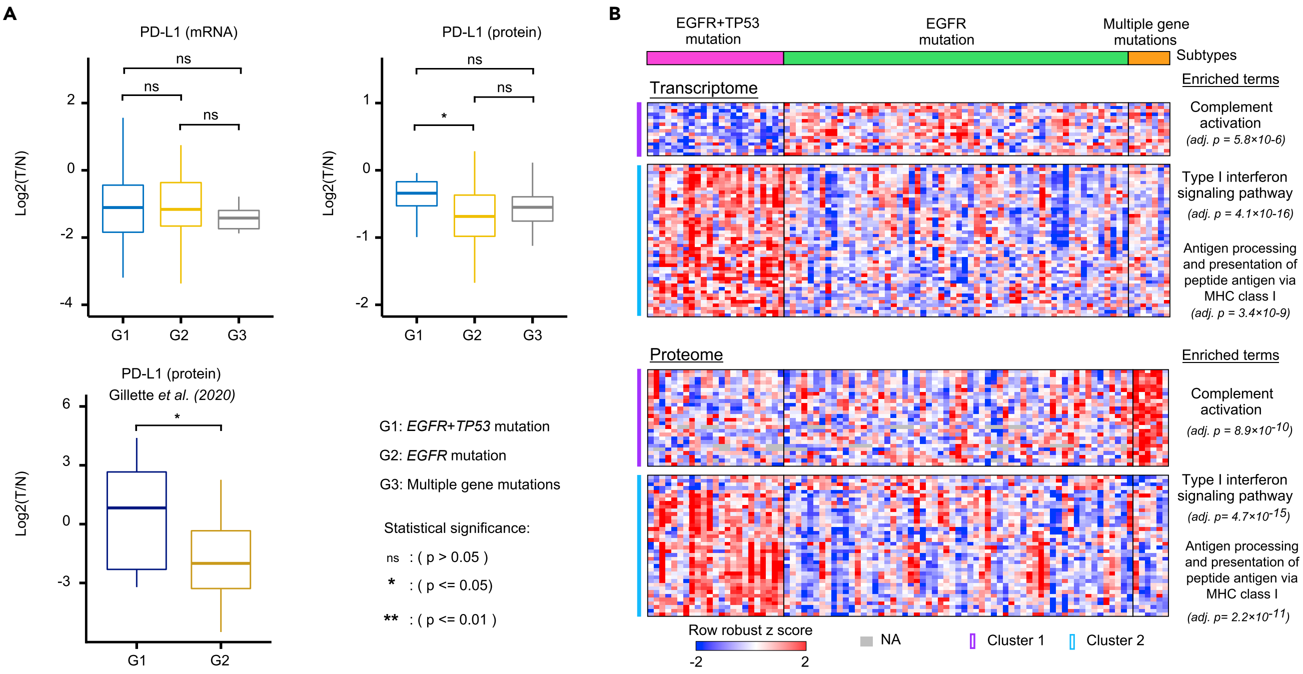


Figure 5. PD-L1 expression and immune characteristics of three subtypes

(A) PD-L1 gene and protein expressions between subtypes in this study (top panel) and the PD-L1 protein expression between G1 and G2 extracted from study of Gillette et al. (bottom panel). Wilcoxon rank-sum test with Bonferroni correction used for statistical comparison between two subtypes.

(B) Immune-related processes revealed by immune-related DE genes/proteins at transcriptomic and proteomic levels. Hierarchical clustering on expression of immune-related DE genes and proteins (shown as heatmap) identifies two clusters of immune-related processes (with Benjamini-Hochberg FDR <0.01). See also Table S5.

Table S5A). These results suggest that patients with the *EGFR+TP53* mutation subtype may have a higher response rate to anti-PD-1/PD-L1 immunotherapy than patients with the *EGFR* mutation subtype.

To explore the immune characteristics between subtypes that could contribute to our understanding of anti-PD-1/PD-L1 therapy responses on the three subtypes, we examined in our cohort the mRNA and protein expressions of 833 immune-related molecules annotated in UniProt (Bateman et al., 2019). A total of 62 immune-related genes (mRNA) and 64 immune-related proteins, respectively, are significantly differentially expressed between at least two of the three subtypes. Hierarchical clustering of the above DE immune-related mRNA and protein expression profiles identified two immune-related clusters, respectively (Tables S5B and S5C), based on which we performed a biological process enrichment analysis and observed that the top enriched terms in the respective clusters at the transcriptomic and proteomic levels could be integrated into two clusters of immune-related processes (Figure 5B). Cluster 1 is related to complement activation (adj. p = 5.8×10^{-6}), in which the mRNA and protein expressions reveal trends that are consistent in G1 and G3, respectively, but inconsistent in G2. Specifically, G1 showed downregulation in the complement activation process, and G3 showed upregulation. Previous reviews and studies report that complement activation in cancer plays both promotion and inhibition (bi-direction) roles in tumor progression (Afshar-Kharghan, 2017; Ajona et al., 2017; Bulla et al., 2016; Pio et al., 2019; Reis et al., 2018). Complement activation probably evokes specific competing pathways in particular cancer subtypes based on the net effect of activation and inhibition (Kleczko et al., 2019). Hence, although the result of Cluster 1 is insufficient for conjecturing the actual role of the complement activation process in subtypes, it does reflect that this process is diverse among subtypes. Cluster 2 is related to the type I interferon signaling pathway (adj. p = 4.7×10^{-15}) and the antigen processing and presentation of peptide antigen via the MHC (major histocompatibility complex) class I process (adj. p = 3.4×10^{-9}). Both mRNA and protein expressions of G1 displayed upregulation on both processes, compared with G2 and G3. Previous studies reported that the type I interferon signaling pathway upregulates PD-L1 expression in various cell types (Bazhin et al., 2018; Budhwani et al., 2018; Jacquelot et al., 2019; Morimoto et al., 2018). Sustained activation of this signaling pathway is involved in the resistance mechanism to immune checkpoint blockade (ICB) therapy, and inhibition of this

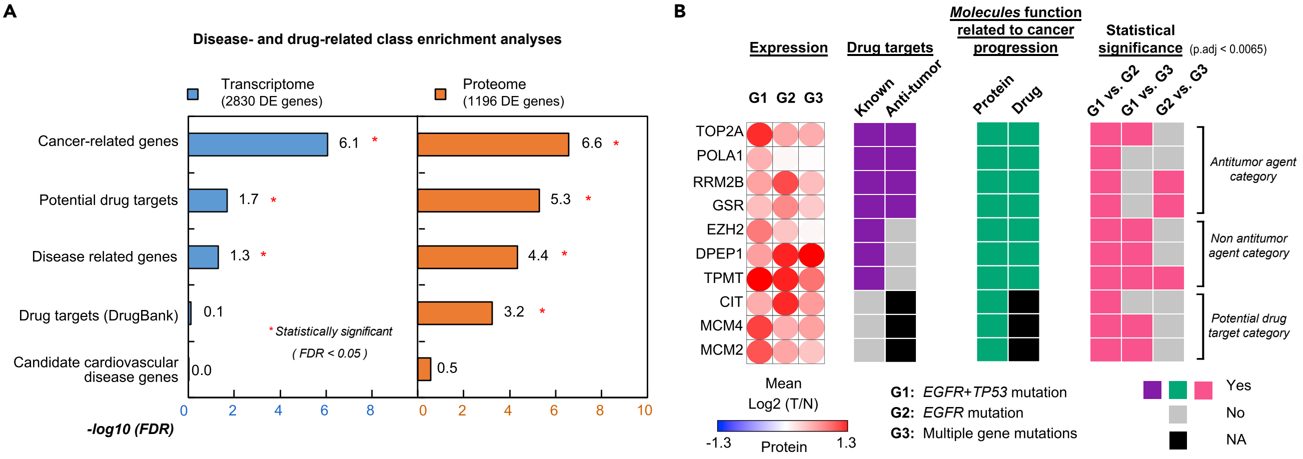


Figure 6. Proteomic analysis between subtypes reveals opportunities for drug repositioning and therapeutics

(A) Comparison of disease- and drug target-related set enrichment analysis on differentially expressed genes (right panel) and proteins (left panel). Enrichment analysis further checked by right-tailed Fisher's exact test and Benjamin-Hochberg procedure for FDR <0.05.

(B) Ten candidate targets for drug repositioning and potential drug targets specific for distinct subtypes. All proteins are related to cancer progression (green color), and seven of them are known drug targets (purple color), including four as anti-tumor drug targets. Statistical significance of pairwise comparison of subtypes (pink color) determined by ANOVA; p < 0.015 and permutation FDR <0.1; Tukey's honest significance test adjusted p < 0.0065. See also Figures S8 and S9 and Table S6.

pathway affects PD-L1 expression and weakens the resistance response of ICB therapy (Benci et al., 2016; Garcia-Diaz et al., 2017). Moreover, tumor cell surface antigens originating from MHC class I antigen processing and presentation (APP) are crucial for CD8+ T-cell recognition and triggering the killing of the target cell (Leone et al., 2013). The defect and downregulation of MHC class I APP in tumor cells both play a role in resistance to ICB therapy (Cai et al., 2018; Gettinger et al., 2017; Sabbatino et al., 2016; Ugurel et al., 2019; Yoo et al., 2019). As observed from Figure 5B, the mRNAs and proteins associated with the processes of Cluster 2 show higher expressions in G1 than in G2 and G3. The upregulated enriched processes of Cluster 2 in G1 suggest that the ICB therapy response of the *EGFR+TP53* mutation subtype is better than that of the other subtypes because of the higher PD-L1 expression induced by the type I interferon signaling pathway and the upregulation of MHC class 1 APP process.

New drug-repositioning and therapeutic opportunities for precision medicine

To compare the druggability of the DE genes/proteins, we performed disease-related and drug target set enrichment analyses, which revealed that the DE genes/proteins are significantly related to the cancer-related gene set, potential drug target gene set, and disease-related gene set at both transcriptomic and proteomic levels (BH FDR < 0.05) (Figure 6A; Table S6A). For the above three enriched gene sets the significance level at the proteomic level is higher than that at the transcriptomic level. In contrast, the DE genes/proteins are not significantly related to the candidate cardiovascular disease gene set, which is used as a negative control set in our analysis. Notably, for the drug target gene set only the DE proteins obtained from the proteomic-level analysis are significantly related (BH FDR=5.95E-04). These results show that the druggability of the DE bio-molecular targets obtained from the proteomic level analysis is superior to that for the DE bio-molecular targets obtained from the transcriptomic level.

We further explored the therapeutic categories of the known drug targets present in the 1,196 DE proteins among the subtypes to discover what kind of disease treatment these proteins are involved in. Among the 33 therapeutic categories obtained from DrugBank (Wishart et al., 2018), only five categories—myelosuppressive agents, antineoplastic agents, cardiotoxic antineoplastic agents, antineoplastic and immunomodulating agents, and immunosuppressive agents—are significantly enriched (BH FDR < 0.05): notably, all cancer-treatment-related categories (Figure S8A; Table S6B). These results reveal that DE proteins are significantly associated with known drug targets for cancer treatments and further imply that DE proteins, which are also known drug targets of the non-cancer-treatment-related class, may play an anti-tumor role in the treatment of specific subtypes. These findings and implications merit further investigation as to the possibility of drug repositioning suitable for the three subtypes.

From the drug target set enrichment analysis, we obtained a total of 76 known drug targets from DE proteins, which were explored to seek the possibility of drug repositioning suitable for the three subtypes (Table S6C). Among these, 30 and 46 drug targets belong to the five significantly enriched cancer-treatment-related categories and 28 non-cancer-treatment-related categories, for convenience termed *anti-tumor agent categories* and *non-anti-tumor agent categories*, respectively. Considering the clinical utility of drug targets with corresponding drugs, we carefully selected seven potential drug targets that satisfy the rigorous criteria among the 76 known drug targets from the DE proteins (STAR Methods, Figures S8B and S8C) for further analysis. First, in the anti-tumor agent categories, we selected the TOP2A (DNA topoisomerase II alpha), POLA1 (DNA polymerase alpha 1, catalytic subunit), RRM2B (Ribonucleotide reductase regulatory TP53 inducible subunit M2B), and GSR (Glutathione-disulfide reductase) proteins as potential drug target candidates for treating specific subtypes (Figure 6B; Table S6D). The protein expressions of TOP2A and POLA1 in G1 are significantly higher than in G2; therefore, the corresponding drugs, such as Etoposide (TOP2A), Idarubicin (TOP2A), Fludarabine (POLA1), and Clofarabine (POLA1), may be more responsive and effective in anticancer activity for the *EGFR*+*TP53* mutation subtype. In contrast, the RRM2B and GSR expressions of G2 are significantly higher than those of G1; thus the corresponding drugs, such as Cladribine (RRM2B) and Carmustine (GSR), may be more sensitive and effective in antitumor activity for the *EGFR* mutation subtype. In summary, these results suggest that different somatic mutation subtypes are responsive to different known anti-tumor agents, and several known anti-tumor drugs can be considered possibly suitable for treatment of different LUAD somatic mutation subtypes.

Next, in the non-anti-tumor agent categories, we selected the following three drug targets—EZH2 (Enhancer of zeste 2 polycomb repressive complex 2 subunit), DPEP1 (Dipeptidase 1), and TPMT (Thiopurine S-methyltransferase) (Figure 6B; Table S6E). The expressions of EZH2 and TPMT are significantly higher in G1 than in G2, whereas the expression of DPEP1 is significantly higher in G2 than in G1. Notably, these three proteins are reported to be involved in processes related to cancer development (Lu et al., 2011; Tiedemann et al., 2012; Toiyama et al., 2011). According to DrugBank (2020-01-02 version), Tazemetostat, Cilastatin, and Olsalazine are known inhibitors of EZH2, DPEP1, and TPMT, respectively, and affect the tumor activity in cancer cell lines, tumor tissues, and xenograft models (Knutson et al., 2013; Park et al., 2016; Velayos et al., 2005) (Table S6E). Specifically, Tazemetostat received FDA approval on January 23, 2020 as a treatment for patients with metastatic or locally advanced epithelioid sarcoma not eligible for complete resection (Leslie, 2020). Cilastatin has been shown to prevent the invasion activity of SW480 cells harboring the DPEP1-expressing vector (Park et al., 2016). Olsalazine has been shown to restrain colorectal cancer progression in patients (Koelink et al., 2010). These findings reveal rising opportunities of new drug repositioning for the treatment of different LUAD somatic mutation subtypes.

To seek new potential drug targets for treating different somatic mutation subtypes, we investigated DE proteins, which were shown earlier to be significantly related to the potential drug target gene set (Benjamin-Hochberg FDR < 0.0001), for possible candidates. Among the 1,196 DE proteins, a total of 124 proteins are annotated as a potential drug target in the HPA database (Table S6F). We selected the following three proteins as high-priority potential drug targets that satisfy our rigorous criteria mentioned in STAR Methods: CIT (citron rho-interacting serine/threonine kinase), MCM2 (minichromosome maintenance complex component 2), and MCM4 (minichromosome maintenance complex component 4) (Figures 6B and S8D; Table S6F). CIT plays a role in the regulation of cytokinesis and cell division (Madaule et al., 1998). Moreover, CIT is reported to be overexpressed in various cancer types, and CIT knockdown reduces cancer cell proliferation (Meng et al., 2019; Sahin et al., 2019; Shou et al., 2020; Wu et al., 2017). Because the expression of the CIT protein in G2 is significantly higher than that in G1, a new drug designed to inhibit CIT activities may favorably impact the cancer therapy progression of the *EGFR* mutation subtype. MCM2 and MCM4 play critical roles in DNA replication initiation and in the elongation process in eukaryotic cells (Forsburg, 2004; Lei, 2005). Overexpression of the MCM complex's subunit is observed in various cancer types (Das et al., 2013; Lau et al., 2010; Wu et al., 2018; Zhong et al., 2017). Notably, the MCM2 and MCM4 proteins have been identified as a promising therapeutic target in human NSCLC treatment, and new compounds have been recently identified to inhibit MCM complex subunits for suppressing tumor cell growth (Byun et al., 2020; Lin et al., 2020). Specifically, a furanonaphthoquinone-based small molecule, AS4583, can bind to the N-terminal domain of MCM2, which significantly decreases the MCM2 level and inhibits tumor growth in AS4583-treated NSCLC cells (Lin et al., 2020). Because MCM2 and MCM4 are more highly expressed in G1 than in G2, theoretically, dosage adjustment of such compounds based on the expression levels of MCM2 and MCM4 in the *EGFR*+*TP53* mutation and *EGFR* mutation subtypes may enhance antitumor activity. In summary, the above bio-molecular level analysis of different LUAD somatic mutation subtypes reveals promising opportunities for drug repositioning and new therapeutic drug targets in precision medicine.

DISCUSSION

In this study, we presented a somatic mutation subtyping analysis of an early stage East Asian LUAD cohort with 88 pairs of tumors and adjacent normal tissues obtained from our recent publication (Chen et al., 2020b), in which we performed an epidemiological analysis and a proteogenomic analysis based on multi-omics profiling. This work is a rational extension arising from therapeutic and molecular biology perspectives to investigate the biological effects of intrinsic somatic mutations in a LUAD cohort. Using robust bioinformatics analysis on the somatic mutation features and multi-omics profiling, we demonstrate that the three somatic mutation subtypes of LUAD exhibit significant differences in some biological processes and in expression levels of known druggable targets, thus possibly affecting the prognosis of cancer. Furthermore, these results provide new insight into the molecular biology of LUAD and reveal new therapeutic opportunities for precision medicine.

Molecular subtyping based on protein or mRNA expressions is usually conducted in large-scale cancer proteogenomic studies to find associations between molecular expression patterns and tumor phenotypes or clinical characteristics (i.e., association between biomolecular expressions and phenotypes) (Johansson et al., 2019; Stewart et al., 2019; Xu et al., 2020; Zhang et al., 2014). However, somatic mutations occurring in DNA could affect gene expression at the transcriptome level and protein functions at the proteome level. In contrast to the aforementioned manner of molecular subtyping, our molecular subtyping is based on somatic mutation genes to investigate the impact of somatic mutation patterns on molecular expressions and clinical outcomes of LUAD (i.e., association of somatic mutation on the biomolecular expressions and phenotypes). Data-driven unsupervised clustering on somatic mutation genes in 88 LUAD samples yielded three subtypes–*EGFR*+*TP53* mutation subtype, *EGFR* mutation subtype, and multiple gene mutation subtype—revealing the diversity and heterogeneity of somatic mutation in the East Asian LUAD cohort. Our results showed that the somatic mutation subtypes revealed diverse protein-mRNA concordance and varying biological processes. Furthermore, we provided an estimation of prognosis for different subtypes and obtained drug-repositioning opportunities and new potential drug targets for different subtypes. However, few large-scale LUAD proteogenomic studies focus on discovering the biological impact of different somatic mutation subtypes. Our findings merit further verification in the future.

Integrating multiple known lung-cancer-related prognostic gene expressions to infer the prognosis, we showed that the *EGFR*+*TP53* mutation subtype has poorer prognosis than other subtypes, which is consistent with studies showing that patients harboring the *EGFR* and *TP53* co-mutation have worse prognosis than patients harboring *EGFR* mutation only (Hou et al., 2019; Jiao et al., 2018; Qin et al., 2020). Notably, our approach for the analysis is quite different from studies that mainly adopt survival analysis. Nevertheless, our results demonstrate the reliability of our approach to infer the prognosis between the cancer subtypes, which can be applied in cancer subtyping studies without patients' survival data in the future.

Analyzing the differentially expressed proteins that harbor drug target annotation in the three subtypes allowed us to propose seven valuable potential drug-repositioning targets and three potential drug targets. To rigorously select drug targets among 200 (30+46+124) DE proteins, we first considered not only the protein expressions but also the druggability of proteins. Next, we investigated whether the proteins and their corresponding drugs are involved in the cancer progression by examining the commercial IPA database and conducting a literature survey. Furthermore, to ensure candidate drug targets are sensitive to different somatic mutation subtypes, we examined the protein expression of the ten selected drug targets and confirmed that they are not significantly different with respect to cancer stages (ANOVA, FDR > 0.2) (Figures S8B–S8D). In other words, the proteins captured by the above criteria and rationale are high-potential cancer-related drug targets derived from somatic mutation genes, which are demonstrated by several studies mentioned in the Result section. However, the reproducibility of potential biomarkers and drug targets identified from a discovery phase study is always a challenge while in the validation and preclinical phase (Baker, 2016; Kaelin, 2017; Prinz et al., 2011). We chose to accomplish this by extracting the expressions of the ten proteins of patients that possess the *EGFR*+*TP53* mutation (n = 21) or *EGFR* mutation only (n = 17) from the recent CPTAC-published large-scale LUAD study (Gillette et al., 2020) to examine the reproducibility of our findings (Table S5A). The results showed that the expression trends of these proteins between the *EGFR* + *TP53* mutation and *EGFR* mutation subtypes are very close to our findings (Figure S9).

Notably, most of these proteins are differentially expressed between the two subtypes. This attests the reproducibility of the evidence for our proposed drug targets for LUAD somatic mutation subtypes in different cohorts and merits further investigation in the future.

In summary, our study is a comprehensive analysis of LUAD somatic mutation subtypes that furthers our understanding of the biological impacts of somatic mutation gene patterns on early stage East Asian patients and provides potential targets for the development of new therapeutic strategies in precision medicine.

Limitations of the study

In this study, based on rigorous analyses we have nominated ten proteins as candidates for drug repositioning and potential drug targets for therapy specific to the somatic mutation subtypes of East Asian LUAD. However, the anti-tumor molecular mechanism of these targets on the somatic mutation subtypes is still unclear. Further research is necessary to investigate and confirm the findings reported hereof to be clinically meaningful.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Acquirement of Taiwan LUAD data
- METHOD DETAILS
  - Mutation features of Taiwan and TCGA LUAD cohorts
  - Somatic mutation subtyping
  - Differential expression analysis
  - Biological process gene set enrichment analysis
  - Protein-mRNA correlation analysis
  - Cancer-related prognostic analysis
  - Disease-related and drug target analysis
  - Drug repositioning and potential drug targets
  - Immune-related gene and proteins analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102522.

ACKNOWLEDGMENTS

This work was supported in part by the Next-Generation Pathway of Taiwan Cancer Precision Medicine Program (Grant No. AS-KPQ-107-TCPMP) at Academia Sinica, Taiwan.

AUTHOR CONTRIBUTIONS

Conceptualization, W.K.C, T.Y.S; Methodology, W.K.C, T.Y.S; Software, W.K.C; Formal Analysis, W.K.C; Investigation, W.K.C, T.Y.S; Computation & statistical analysis, W.K.C; Resources, W.K.C T.Y.S; Data Curation, W.K.C; Writing—Review & Editing, T.Y.S, W.K.C; Visualization, W.K.C, T.Y.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 23, 2020
Revised: March 25, 2021
Accepted: May 4, 2021
Published: June 25, 2021

REFERENCES

Afshar-Kharghan, V. (2017). The role of the complement system in cancer. J. Clin. Invest. *127*, 780–789.

Ajona, D., Ortiz-Espinosa, S., Moreno, H., Lozano, T., Pajares, M.J., Agorreta, J., Bertolo, C., Lasarte, J.J., Vicent, S., Hoehlig, K., et al. (2017). A combined PD-1/C5a blockade synergistically protects against lung cancer growth and metastasis. Cancer Discov. *7*, 694–703.

Arbour, K.C., Jordan, E., Kim, H.R., Dienstag, J., Yu, H.A., Sanchez-Vega, F., Lito, P., Berger, M., Solit, D.B., Hellmann, M., et al. (2018). Effects of co-occurring genomic alterations on outcomes in patients with KRAS-mutant non-small cell lung cancer. Clin. Cancer Res. *24*, 334–340.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature *533*, 452–454.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Alpi, E., Bely, B., Bingley, M., Britto, R., Bursteinas, B., Busiello, G., et al. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Bazhin, A.V., von Ahn, K., Fritz, J., Werner, J., and Karakhanova, S. (2018). Interferon-alpha up-regulates the expression of PD-L1 molecules on immune cells through STAT3 and p38 signaling. Front. Immunol. *9*, 2129.

Benci, J.L., Xu, B., Qiu, Y., Wu, T.J., Dada, H., Twyman-Saint Victor, C., Cucolo, L., Lee, D.S.M., Pauken, K.E., Huang, A.C., et al. (2016). Tumor interferon signaling regulates a multigenic resistance program to immune checkpoint blockade. Cell *167*, 1540–1554.

Bhattacharya, B., Mohd Omar, M.F., and Soong, R. (2016). The Warburg effect and drug resistance. Br. J. Pharmacol. *173*, 970–979.

BITGDACenter. (2016). Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run, *10* (Broad Institute of MIT and Harvard Dataset), p. C11G0KM9.

Blakely, C.M., Watkins, T.B.K., Wu, W., Gini, B., Chabon, J.J., McCoach, C.E., McGranahan, N., Wilson, G.A., Birkbak, N.J., Olivas, V.R., et al. (2017). Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. Nat. Genet. *49*, 1693–1704.

Bolstad, B. (2019). preprocessCore: A Collection of Pre-processing Functions 1.46. 0, R. Package Version. 10.18129/B9.bioc.preprocessCore

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Budhwani, M., Mazzieri, R., and Dolcetti, R. (2018). Plasticity of type I interferon-mediated responses in cancer therapy: from anti-tumor immunity to resistance. Front. Oncol. *8*, 322.

Bulla, R., Tripodo, C., Rami, D., Ling, G.S., Agostinis, C., Guarnotta, C., Zorzet, S., Durigutto, P., Botto, M., and Tedesco, F. (2016). C1q acts in the tumour microenvironment as a cancer-promoting factor independently of complement activation. Nat. Commun. *7*, 1–11.

Byun, W.S., Kim, S., Shin, Y.H., Kim, W.K., Oh, D.C., and Lee, S.K. (2020). Antitumor activity of Ohmyungsamycin A through the regulation of the Skp2-p27 axis and MCM4 in human colorectal cancer cells. J. Nat. Prod. *83*, 118–126.

Cai, L., Michelakos, T., Yamada, T., Fan, S., Wang, X., Schwab, J.H., Ferrone, C.R., and Ferrone, S. (2018). Defective HLA class I antigen processing machinery in cancer. Cancer Immunol. Immunother. *67*, 999–1009.

Canale, M., Petracci, E., Delmonte, A., Chiadini, E., Dazzi, C., Papi, M., Capelli, L., Casanova, C., De Luigi, N., Mariotti, M., et al. (2017). Impact of TP53 mutations on outcome in EGFR-mutated patients treated with first-line tyrosine kinase inhibitors. Clin. Cancer Res. *23*, 2195–2202.

CancerGenomeAtlasResearchNetwork (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Chen, F., Zhang, Y., Parra, E., Rodriguez, J., Behrens, C., Akbani, R., Lu, Y., Kurie, J.M., Gibbons, D.L., Mills, G.B., et al. (2017). Multiplatform-based molecular subtypes of non-small-cell lung cancer. Oncogene *36*, 1384–1393.

Chen, J., Yang, H., Teo, A.S.M., Amer, L.B., Sherbaf, F.G., Tan, C.Q., Alvarez, J.J.S., Lu, B., Lim, J.Q., Takano, A., et al. (2020a). Genomic landscape of lung adenocarcinoma in East Asians. Nat. Genet. *52*, 177–186.

Chen, Y.J., Roumeliotis, T.I., Chang, Y.H., Chen, C.T., Han, C.L., Lin, M.H., Chen, H.W., Chang, G.C., Chang, Y.L., Wu, C.T., et al. (2020b). Proteogenomics of non-smoking lung cancer in east assia delineates molecular signatures of pathogenesis and progression. Cell *182*, 226–244.

Chen, Z., Fillmore, C.M., Hammerman, P.S., Kim, C.F., and Wong, K.K. (2014). Non-small-cell lung cancers: a heterogeneous set of diseases. Nat. Rev. Cancer *14*, 535–546.

Das, M., Prasad, S.B., Yadav, S.S., Govardhan, H.B., Pandey, L.K., Singh, S., Pradhan, S., and Narayan, G. (2013). Over expression of minichromosome maintenance genes is clinically correlated to cervical carcinogenesis. PLoS One *8*, e69607.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*, 1069–1075.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The reactome pathway knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Fehrenbacher, L., Spira, A., Ballinger, M., Kowanetz, M., Vansteenkiste, J., Mazieres, J., Park, K., Smith, D., Artal-Cortes, A., Lewanski, C., et al. (2016). Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. Lancet *387*, 1837–1846.

Forsburg, S.L. (2004). Eukaryotic MCM proteins: beyond replication initiation. Microbiol. Mol. Biol. Rev. *68*, 109–131.

Gahr, S., Stoehr, R., Geissinger, E., Ficker, J.H., Brueckl, W.M., Gschwendtner, A., Gattenloehner, S., Fuchs, F.S., Schulz, C., Rieker, R.J., et al. (2013). EGFR mutational status in a large series of Caucasian European NSCLC patients: data from daily practice. Br. J. Cancer *109*, 1821–1828.

Garcia-Diaz, A., Shin, D.S., Moreno, B.H., Saco, J., Escuin-Ordinas, H., Rodriguez, G.A., Zaretsky, J.M., Sun, L., Hugo, W., Wang, X., et al. (2017). Interferon receptor signaling pathways regulating PD-L1 and PD-L2 expression. Cell Rep. *19*, 1189–1201.

Gettinger, S., Choi, J., Hastings, K., Truini, A., Datar, I., Sowell, R., Wurtz, A., Dong, W., Cai, G., Melnick, M.A., et al. (2017). Impaired HLA class I antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. Cancer Discov. *7*, 1420–1435.

Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *182*, 200–225.

Hennig, C. (2020). fpc: Flexible Procedures for Clustering. R package version 2.2-5, https://CRAN.R-project.org/package=fpc.

Herbst, R.S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature *553*, 446–454.

Hou, H., Qin, K., Liang, Y., Zhang, C., Liu, D., Jiang, H., Liu, K., Zhu, J., Lv, H., Li, T., et al. (2019). Concurrent TP53 mutations predict poor outcomes of EGFR-TKI treatments in Chinese patients with advanced NSCLC. Cancer Manag. Res. *11*, 5665–5675.

Hua, C., Zhao, G., Li, Y., and Bie, L. (2014). Minichromosome maintenance (MCM) family as potential diagnostic and prognostic tumor markers for human gliomas. BMC Cancer *14*, 1–8.

Icard, P., Shulman, S., Farhat, D., Steyaert, J.M., Alifano, M., and Lincet, H. (2018). How the Warburg effect supports aggressiveness and drug resistance of cancer cells? Drug Resist. Updat. *38*, 1–11.

Jacquelot, N., Yamazaki, T., Roberti, M.P., Duong, C.P.M., Andrews, M.C., Verlingue, L., Ferrere, G., Becharef, S., Vetizou, M., Daillere, R., et al. (2019). Sustained type I interferon signaling as a mechanism of resistance to PD-1 blockade. Cell Res. *29*, 846–861.

Jiao, X.D., Qin, B.D., You, P., Cai, J., and Zang, Y.S. (2018). The prognostic value of TP53 and its correlation with EGFR mutation in advanced non-small cell lung cancer, an analysis based on cBioPortal data base. Lung Cancer *123*, 70–75.

Johansson, H.J., Socciarelli, F., Vacanti, N.M., Haugen, M.H., Zhu, Y., Siavelis, I., Fernandez-Woodbridge, A., Aure, M.R., Sennblad, B., Vesterlund, M., et al. (2019). Breast cancer quantitative proteome and proteogenomic landscape. Nat. Commun. *10*, 1–14.

Jordan, E.J., Kim, H.R., Arcila, M.E., Barron, D., Chakravarty, D., Gao, J., Chang, M.T., Ni, A., Kundra, R., Jonsson, P., et al. (2017). Prospective comprehensive molecular characterization of lung adenocarcinomas for efficient patient matching to approved and emerging therapies. Cancer Discov. *7*, 596–609.

Kaelin, W.G., Jr. (2017). Common pitfalls in preclinical cancer target validation. Nat. Rev. Cancer *17*, 425–440.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kleczko, E.K., Kwak, J.W., Schenk, E.L., and Nemenoff, R.A. (2019). Targeting the complement pathway as a therapeutic strategy in lung cancer. Front. Immunol. *10*, 954.

Knutson, S.K., Warholic, N.M., Wigle, T.J., Klaus, C.R., Allain, C.J., Raimondi, A., Porter Scott, M., Chesworth, R., Moyer, M.P., Copeland, R.A., et al. (2013). Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase EZH2. Proc. Natl. Acad. Sci. U S A *110*, 7922–7927.

Koelink, P.J., Hawinkels, L.J., Wiercinska, E., Sier, C.F., ten Dijke, P., Lamers, C.B., Hommes, D.W., and Verspaget, H.W. (2010). 5-Aminosalicylic acid inhibits TGF-beta1 signalling in colorectal cancer cells. Cancer Lett. *287*, 82–90.

Labbe, C., Cabanero, M., Korpanty, G.J., Tomasini, P., Doherty, M.K., Mascaux, C., Jao, K., Pitcher, B., Wang, R., Pintilie, M., et al. (2017). Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC). Lung Cancer *111*, 23–29.

Lau, K.M., Chan, Q.K., Pang, J.C., Li, K.K., Yeung, W.W., Chung, N.Y., Lui, P.C., Tam, Y.S., Li, H.M., Zhou, L., et al. (2010). Minichromosome maintenance proteins 2, 3 and 7 in medulloblastoma: overexpression and involvement in regulation of cell migration and invasion. Oncogene *29*, 5475–5489.

Lei, M. (2005). The MCM complex: its role in DNA replication and implications for cancer therapy. Curr. Cancer Drug Targets *5*, 365–380.

Leone, P., Shin, E.C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. J. Natl. Cancer Inst. *105*, 1172–1187.

Leslie, M. (2020). First EZH2 inhibitor approved-for rare sarcoma. Cancer Discov. *10*, 333–334.

Liao, Y.X., Wang, J., Jaehnig, E.J., Shi, Z.A., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205.

Lin, C.Y., Wu, H.Y., Hsu, Y.L., Cheng, T.R., Liu, J.H., Huang, R.J., Hsiao, T.H., Wang, C.J., Hung, P.F., Lan, A., et al. (2020). Suppression of drug-resistant non-small-cell lung cancer with inhibitors targeting minichromosomal maintenance protein. J. Med. Chem. *63*, 3172–3187.

Liu, Z., Li, J., Chen, J., Shan, Q., Dai, H., Xie, H., Zhou, L., Xu, X., and Zheng, S. (2018). MCM family in HCC: MCM6 indicates adverse tumor features and poor outcomes and promotes S/G2 cell cycle progression. BMC Cancer *18*, 1–10.

Lu, J., He, M.L., Wang, L., Chen, Y., Liu, X., Dong, Q., Chen, Y.C., Peng, Y., Yao, K.T., Kung, H.F., et al. (2011). MiR-26a inhibits cell growth and tumorigenesis of nasopharyngeal carcinoma through repression of EZH2. Cancer Res. *71*, 225–233.

Lunt, S.Y., and Vander Heiden, M.G. (2011). Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. Annu. Rev. Cell Dev. Biol. *27*, 441–464.

Lv, Z., and Lei, T. (2020). Systematical identifications of prognostic meaningful lung adenocarcinoma subtypes and the underlying mutational and expressional characters. BMC Cancer *20*, 56.

Madaule, P., Eda, M., Watanabe, N., Fujisawa, K., Matsuoka, T., Bito, H., Ishizaki, T., and Narumiya, S. (1998). Role of citron kinase as a target of the small GTPase Rho in cytokinesis. Nature *394*, 491–494.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). cluster: Cluster analysis basics and extensions. R package version 2.1.0, https://CRAN.R-project.org/package=cluster.

Mellman, I., Coukos, G., and Dranoff, G. (2011). Cancer immunotherapy comes of age. Nature *480*, 480–489.

Meng, D., Yu, Q., Feng, L., Luo, M., Shao, S., Huang, S., Wang, G., Jing, X., Tong, Z., Zhao, X., et al. (2019). Citron kinase (CIT-K) promotes aggressiveness and tumorigenesis of breast cancer cells in vitro and in vivo: preliminary study of the underlying mechanism. Clin. Transl. Oncol. *21*, 910–923.

Morimoto, Y., Kishida, T., Kotani, S.I., Takayama, K., and Mazda, O. (2018). Interferon-beta signal may up-regulate PD-L1 expression through IRF9-dependent and independent pathways in lung cancer cells. Biochem. Biophys. Res. Commun. *507*, 330–336.

Nahar, R., Zhai, W., Zhang, T., Takano, A., Khng, A.J., Lee, Y.Y., Liu, X., Lim, C.H., Koh, T.P.T., Aung, Z.W., et al. (2018). Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing. Nat. Commun. *9*, 1–11.

Park, S.Y., Lee, S.J., Cho, H.J., Kim, T.W., Kim, J.T., Kim, J.W., Lee, C.H., Kim, B.Y., Yeom, Y.I., Lim, J.S., et al. (2016). Dehydropeptidase 1 promotes metastasis through regulation of E-cadherin expression in colon cancer. Oncotarget *7*, 9501–9512.

Pio, R., Ajona, D., Ortiz-Espinosa, S., Mantovani, A., and Lambris, J.D. (2019). Complementing the cancer-immunity cycle. Front. Immunol. *10*, 774.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. *10*, 712.

Qin, K., Hou, H., Liang, Y., and Zhang, X. (2020). Prognostic value of TP53 concurrent mutations for EGFR- TKIs and ALK-TKIs based targeted therapy in advanced non-small cell lung cancer: a meta-analysis. BMC Cancer *20*, 1–16.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198.

RCoreTeam (2019). R: A Language and Environment for Statistical Computing, https://www.R-project.org.

Reck, M., Rodriguez-Abreu, D., Robinson, A.G., Hui, R., Csoszi, T., Fulop, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., et al. (2016). Pembrolizumab versus Chemotherapy for PD-L1-positive non-small-cell lung cancer. N. Engl. J. Med. *375*, 1823–1833.

Reis, E.S., Mastellos, D.C., Ricklin, D., Mantovani, A., and Lambris, J.D. (2018). Complement in cancer: untangling an intricate relationship. Nat. Rev. Immunol. *18*, 1–5.

Rosell, R., Carcereny, E., Gervais, R., Vergnenegre, A., Massuti, B., Felip, E., Palmero, R., Garcia-Gomez, R., Pallares, C., Sanchez, J.M., et al. (2012). Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. Lancet Oncol. *13*, 239–246.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Sabbatino, F., Villani, V., Yearley, J.H., Deshpande, V., Cai, L., Konstantinidis, I.T., Moon, C., Nota, S., Wang, Y., Al-Sukaini, A., et al. (2016). PD-L1 and HLA class I antigen expression and clinical course of the disease in intrahepatic cholangiocarcinoma. Clin. Cancer Res. *22*, 470–478.

Sahin, I., Kawano, Y., Sklavenitis-Pistofidis, R., Moschetta, M., Mishima, Y., Manier, S., Sacco, A., Carrasco, R., Fonseca, R., Roccaro, A.M., et al. (2019). Citron Rho-interacting kinase silencing causes cytokinesis failure and reduces tumor growth in multiple myeloma. Blood Adv. *3*, 995–1002.

Scheffler, M., Ihle, M.A., Hein, R., Merkelbach-Bruse, S., Scheel, A.H., Siemanowski, J., Bragelmann, J., Kron, A., Abedpour, N., Ueckeroth, F., et al. (2019). K-ras mutation subtypes in NSCLC and associated co-occuring mutations in other oncogenic pathways. J. Thorac. Oncol. *14*, 606–616.

Shigematsu, H., Lin, L., Takahashi, T., Nomura, M., Suzuki, M., Wistuba, I.I., Fong, K.M., Lee, H., Toyooka, S., Shimizu, N., et al. (2005). Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. J. Natl. Cancer Inst. *97*, 339–346.

Shou, J., Yu, C., Zhang, D., and Zhang, Q. (2020). Overexpression of citron rho-interacting serine/threonine kinase associated with poor outcome in bladder cancer. J. Cancer *11*, 4173–4180.

Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA Cancer J. Clin. *69*, 7–34.

Skoulidis, F., Byers, L.A., Diao, L., Papadimitrakopoulou, V.A., Tong, P., Izzo, J., Behrens, C., Kadara, H., Parra, E.R., Canales, J.R., et al. (2015). Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov. *5*, 860–877.

Skoulidis, F., and Heymach, J.V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. Nat. Rev. Cancer *19*, 495–509.

Solomon, B.J., Mok, T., Kim, D.W., Wu, Y.L., Nakagawa, K., Mekhail, T., Felip, E., Cappuzzo, F., Paolini, J., Usari, T., et al. (2014). First-line crizotinib versus chemotherapy in ALK-positive lung cancer. N. Engl. J. Med. *371*, 2167–2177.

Stewart, P.A., Welsh, E.A., Slebos, R.J.C., Fang, B., Izumi, V., Chambers, M., Zhang, G., Cen, L., Pettersson, F., Zhang, Y., et al. (2019). Proteogenomic landscape of squamous cell lung cancer. Nat. Commun. *10*, 1–17.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.

Tiedemann, R.E., Zhu, Y.X., Schmidt, J., Shi, C.X., Sereduk, C., Yin, H.W., Mousses, S., and Stewart, A.K. (2012). Identification of molecular vulnerabilities in human multiple myeloma cells by RNA interference lethality screening of the druggable genome. Cancer Res. *72*, 757–768.

Toiyama, Y., Inoue, Y., Yasuda, H., Saigusa, S., Yokoe, T., Okugawa, Y., Tanaka, K., Miki, C., and Kusunoki, M. (2011). DPEP1, expressed in the early stages of colon carcinogenesis, affects cancer cell invasiveness. J. Gastroenterol. *46*, 153–163.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U S A *98*, 5116–5121.

Tyanova, S., and Cox, J. (2018). Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research. Methods Mol. Biol. *1711*, 133–148.

Ugurel, S., Spassova, I., Wohlfarth, J., Drusio, C., Cherouny, A., Melior, A., Sucker, A., Zimmer, L., Ritter, C., Schadendorf, D., et al. (2019). MHC class-I downregulation in PD-1/PD-L1 inhibitor refractory Merkel cell carcinoma and its potential reversal by histone deacetylase inhibition: a case series. Cancer Immunol. Immunother. *68*, 983–990.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based human protein atlas. Nat. Biotechnol. *28*, 1248–1250.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029–1033.

Velayos, F.S., Terdiman, J.P., and Walsh, J.M. (2005). Effect of 5-aminosalicylate use on colorectal cancer and dysplasia risk: a systematic review and metaanalysis of observational studies. Am. J. Gastroenterol. *100*, 1345–1353.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

Wu, C.C., Hsu, H.Y., Liu, H.P., Chang, J.W., Chen, Y.T., Hsieh, W.Y., Hsieh, J.J., Hsieh, M.S., Chen, Y.R., and Huang, S.F. (2008). Reversed mutation rates of KRAS and EGFR genes in adenocarcinoma of the lung in Taiwan and their implications. Cancer *113*, 3199–3208.

Wu, W., Wang, X., Shan, C., Li, Y., and Li, F. (2018). Minichromosome maintenance protein 2 correlates with the malignant status and regulates proliferation and cell cycle in lung squamous cell carcinoma. Oncotargets Ther. *11*, 5025–5034.

Wu, Z., Zhu, X., Xu, W., Zhang, Y., Chen, L., Qiu, F., Zhang, B., Wu, L., Peng, Z., and Tang, H. (2017). Up-regulation of CIT promotes the growth of colon cancer cells. Oncotarget *8*, 71954–71964.

Xu, J.Y., Zhang, C., Wang, X., Zhai, L., Ma, Y., Mao, Y., Qian, K., Sun, C., Liu, Z., Jiang, S., et al. (2020). Integrative proteomic characterization of human lung adenocarcinoma. Cell *182*, 245–261.

Yang, C.Y., Yang, J.C., and Yang, P.C. (2020). Precision management of advanced non-small cell lung cancer. Annu. Rev. Med. *71*, 117–136.

Yang, J.C., Wu, Y.L., Schuler, M., Sebastian, M., Popat, S., Yamamoto, N., Zhou, C., Hu, C.P., O'Byrne, K., Feng, J., et al. (2015). Afatinib versus cisplatin-based chemotherapy for EGFR mutation-positive lung adenocarcinoma (LUX-Lung 3 and LUX-Lung 6): analysis of overall survival data from two randomised, phase 3 trials. Lancet Oncol. *16*, 141–151.

Yoo, S.H., Keam, B., Ock, C.Y., Kim, S., Han, B., Kim, J.W., Lee, K.W., Jeon, Y.K., Jung, K.C., Chung, E.J., et al. (2019). Prognostic value of the association between MHC class I downregulation and PD-L1 upregulation in head and neck squamous cell carcinoma patients. Sci. Rep. *9*, 1–9.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhang, X.C., Wang, J., Shao, G.G., Wang, Q., Qu, X., Wang, B., Moy, C., Fan, Y., Albertyn, Z., Huang, X., et al. (2019). Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients. Nat. Commun. *10*, 1–12.

Zhong, H., Chen, B., Neves, H., Xing, J., Ye, Y., Lin, Y., Zhuang, G., Zhang, S.D., Huang, J., and Kwok, H.F. (2017). Expression of minichromosome maintenance genes in renal cell carcinoma. Cancer Manag. Res. *9*, 637–647.

Zhou, C., Wu, Y.L., Chen, G., Feng, J., Liu, X.Q., Wang, C., Zhang, S., Wang, J., Zhou, S., Ren, S., et al. (2011). Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. Lancet Oncol. *12*, 735–742.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1–10.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Tumor and adjacent normal tissues from 88 Taiwan LUAD patients	Chen et al., 2020b	https://doi.org/10.1016/j.cell.2020.06.012
Tumor and adjacent normal tissues from 38 LUAD patients with EGFR+TP53 comutation or EGFR mutation	Gillette et al., 2020	https://doi.org/10.1016/j.cell.2020.06.013
TCGA lung adenocarcinoma	Broad Institute, FireBrowse Data Portal	http://firebrowse.org
Human Protein Atlas (HPA)	Uhlen et al., 2010	https://www.proteinatlas.org
DrugBank	Wishart et al., 2018	https://go.drugbank.com
UniProt	Bateman et al., 2019	https://www.uniprot.org
Software and algorithms		
cluster (R package)	Maechler et al., 2019	https://cran.r-project.org/web/packages/cluster/index.html
ConsensusClusterPlus (R package)	Wilkerson and Hayes, 2010	https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
fpc (R package)	Hennig, 2020	https://cran.r-project.org/web/packages/fpc/index.html
Metascape	Zhou et al., 2019	http://metascape.org
base (R package)	RCoreTeam, 2019	https://www.r-project.org
WebGestalt	Liao et al., 2019	http://www.webgestalt.org
Morpheus	Broad Institute	https://software.broadinstitute.org/morpheus
g:Profiler	Raudvere et al., 2019	https://biit.cs.ut.ee/gprofiler
stats (R package)	RCoreTeam, 2019	https://www.r-project.org
preprocessCore (R package)	Bolstad, 2019	https://github.com/bmbolstad/preprocessCore
Ingenuity Pathway Analysis (IPA®)	Qiagen	http://www.qiagen.com
Custom R code for this study	This study	https://github.com/ComicsLab/LUAD_SomaticMutationSubtyping

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ting-Yi Sung (tsung@iis.sinica.edu.tw).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The mutation, clinical, proteomic, and transcriptomic data used in this study can be found in Table S1 of Chen et al. (https://doi.org/10.1016/j.cell.2020.06.012). Code used for the analysis in this study is available on Github (https://github.com/ComicsLab/LUAD_SomaticMutationSubtyping).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Acquirement of Taiwan LUAD data

The clinical data, transcriptome profiling, proteome profiling, and somatic mutation features of paired tumor and adjacent normal tissues of 88 treatment-naïve LUAD patients used in this study were obtained from our previous study by Chen et al. (Chen et al., 2020b), in which the data of 103 NSCLC patients were collected and studied. To maintain homogeneity among NSCLC samples for investigating somatic mutation subtyping, we selected the predominant 88 adenocarcinoma (ADC) patients and filtered out 15 squamous cell carcinoma (SCC) or other patients in this study because ADC and SCC are well known to be distinct subtypes of NSCLC and only patients having mutation, transcriptomic, and proteomic data were used. For the 88 patient samples, we retrieved their clinical data, somatic mutation information, transcriptome profiling, and proteome profiling from Tables S1A, S1C, S1D, and S1E in Chen et al., respectively. Moreover, we applied quantile normalization (preprocessCore R package (Bolstad, 2019)) on the retrieved gene and protein expression data (log2 T/N, with T: abundance in tumor and N: abundance in normal tissue) for downstream data analysis in this study. A total of 30,155 RNAs and 13,457 unique proteins were quantified. Detailed information on the sample collection, sample preparation, experimental preparation, genomic and transcriptomic analysis, and mass-spectrometry-based proteomic analysis are provided in Chen et al.

METHOD DETAILS

Mutation features of Taiwan and TCGA LUAD cohorts

The somatic mutation features of a TCGA LUAD cohort (*n* = 574, TCGA data version 2016_01_28) were obtained from the Broad Institute FireBrowse Data Portal via RESTful API interfaces (http://firebrowse.org/). We applied Fisher's exact test on each of the 10 mutation genes with top sample frequencies in our cohort to compare their frequency with that in the TCGA cohort to assess the occurrence differences between the two cohorts. The p value was further adjusted using the Benjamini-Hochberg method (adj. $p < 0.05$) for multiple testing correction.

Somatic mutation subtyping

The Taiwan 88 LUAD samples contain 10,054 somatic mutation genes, each of which includes at least one variant event, such as nonsynonymous SNV (single-nucleotide variant), non-frameshift INDEL, frameshift INDEL, stop-gain, and stop-loss. We converted the somatic mutation information into a 10054x88 binary matrix, referred to as a *somatic mutation gene matrix*, with an entry $m_{ij} = 1$ if at least one variant event occurs in the *i*-th somatic mutation gene of the *j*-th sample and $m_{ij} = 0$ otherwise. Note that to maintain the importance of individual somatic mutation genes in the cohort, we sorted the somatic mutation genes of the binary matrix in decreasing order of their occurrences in the cohort, i.e., the gene with the highest occurrence was the first gene in the matrix. Moreover, we used the cluster R package (Maechler et al., 2019) to convert the somatic mutation gene matrix into a pairwise dissimilarity matrix of Gower's distance and subsequently used the ConsensusClusterPlus R package (Wilkerson and Hayes, 2010) to perform unsupervised consensus clustering based on partitioning around medoids (PAM), with 1,000 iterations, setting maxK to 10 and the other parameters to the default. To explore the number of somatic mutation genes to be included in the somatic mutation gene matrix to ensure optimally efficient clustering, we adopted a ''decrease progressively'' manner to discard unnecessary somatic mutation genes. To be specific, we generated the matrices with all (10,054 genes), the top 5,000, 2,500, 1,200, 600, 300, 150, 100, 50, 25, 15, and 5 somatic mutation genes, respectively, for consensus clustering analysis. The consensus cumulative distribution function plot and delta area plot of the clustering results on the above binary matrices show that optimally efficient clustering appears when using the top 5 to 50 somatic mutation genes, and the best cluster number (k) is mostly 3 or 4 (Figure S1B). Then we examined the results of clustering into three and four groups (k = 3, 4) using different somatic mutation gene matrices of top *n* genes with *n*: 5–50 (Figure S1C). The figure shows that the somatic mutation gene matrix generated by top 20 genes provides stable and optimally efficient clustering. Next, to determine the best number of clusters (k) using the top 20 somatic mutation gene matrix, we generated the discriminant projection plot using the fpc R package (Hennig, 2020) and the silhouette plot using the cluster R package (Figures S1D and S1E). The above two plots revealed that the best number of clusters is three (k = 3) and the average silhouette width is 0.27. Specifically, the three clusters represent three different subtypes, i.e., G1 of 23 patients and evident in co-mutation of EGFR and TP53, G2 of 58 patients and evident in EGFR mutation only, and G3 of 7 patients and evident in multiple gene mutations (Figure 1B; Table S1C). The variety of EGFR mutation in the three subtypes is summarized in Figure S2A. The largest variety of EGFR mutation is Exon19del (39% of 23 patients) in G1, L858R (40% of 58 patients) in G2, and Exon19del and L858R (43% of 7 patients, respectively) in G3. However, the frequency of EGFR mutation varieties is not significantly different between G1 and G2 (Figure S2B). To justify our clustering results, we also performed other two consensus clustering algorithms–consensus K-means and consensus hierarchical clustering–on the top 20 somatic mutation gene matrix, same as the consensus PAM algorithm. Both consensus K-means and consensus hierarchical clustering mainly cluster patients with EGFR mutation only and co-mutated EGFR and TP53 into one dominating group, as shown in Figure S2C. Thus our consensus PAM clustering is more reasonable than the other two consensus clustering algorithms.

Differential expression analysis

To further validate the soundness of the above subtyping based on somatic mutation genes, we first generated three clinical characteristic group datasets and four random sampling group datasets. In the clinical characteristic group datasets, patients were clustered based on age (2 groups: age $\leq$ 60 years and age >60 years), gender (2 groups: male and female), and stage (4 groups: IA, IB, II, and late stage: III and IV). In the random sampling group datasets, four datasets were generated by randomly sampling patients into different numbers of groups: two groups (44 patients each), three groups (30, 30, and 28 patients, respectively), four groups (22 patients each), and five groups (18, 18, 18, 18, and 17 patients, respectively). We individually performed differential expression analyses at transcriptomic and proteomic levels based on log2 T/N values on each of the above eight grouping of patients and the clustering of the three somatic mutation subtypes for comparison. The ANOVA (analysis of variance) test (for >2 groups) or t test (for 2 groups) and permutation FDR correction (Tusher et al., 2001; Tyanova and Cox, 2018) (number of randomizations: 300) were adopted to determine differentially expressed genes/proteins among *k* groups with ANOVA at $p < 0.05$ and FDR <0.1. Moreover, we further performed the Tukey HSD (Honestly Significant Difference) test on the above differentially expressed genes/proteins to determine whether each was statistically significant between any two somatic mutation subtypes (adj. $p < 0.05$). Note that if a gene/protein had a missing value (NA) for more than 50% of the patients in any group, it was excluded from the ANOVA and Tukey tests. The genes/proteins passing the Tukey test were regarded as statistically differentially expressed (abbreviated as DE) genes/proteins. All statistical tests were carried out using the R Stats package (RCoreTeam, 2019).

Biological process gene set enrichment analysis

The pathway and biological process enrichment analysis of the DE genes/proteins between any two somatic mutation subtypes was analyzed using the Metascape platform (http://metascape.org) (Zhou et al., 2019). The KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) pathways, canonical pathways (Subramanian et al., 2005), GO (Gene Ontology) biological processes (Ashburner et al., 2000), Reactome (Fabregat et al., 2018), and CORUM (comprehensive resource of mammalian protein complexes) (Ruepp et al., 2010) gene sets were selected for enrichment analysis. An FDR value (Benjamini-Hochberg procedure) of 0.05 and at least three genes/proteins per term were used as the cutoff to define statistically significant enriched terms.

Protein-mRNA correlation analysis

A total of 30,155 mRNAs and 13,457 proteins were quantified in the 88 LUAD samples, respectively, among which a total of 9,009 genes were found expressed in both mRNA and protein levels of a majority of patients, satisfying that each gene had a missing value in <50% of the patents in each subtype. These 9,009 genes were used to calculate the sample-wise and gene-wise protein-mRNA correlation by the Spearman correlation coefficients (rho) using the base R package (RCoreTeam, 2019). The gene-wise protein-mRNA rho values of each somatic mutation subtype were used for the KEGG pathway enrichment analysis conducted on the WebGestalt server. The parameters for the enrichment analysis were 25 for the minimum and 150 for the maximum numbers of IDs in each category, 3,000 for the number of permutations; and FDR <0.05 for the significance level. The Kolmogorov-Smirnov test with Benjamini-Hochberg adjustment (adj. $p < 0.05$) was applied to assess whether the enriched KEGG pathways had different distributions of the protein-mRNA correlations between groups.

Cancer-related prognostic analysis

Cancer-related prognostic genes were obtained from the HPA database (version: 19.3), which provides prognostic genes with favorable and unfavorable clinical outcomes for each of 17 cancer types. Among a total of 10,491 prognostic genes recorded in HPA, 6,771 genes and 6,055 genes were annotated as unfavorable and favorable prognostic genes for different cancer types, respectively. Prognostic gene set enrichment analysis on the DE genes was performed with Pearson's chi-squared test. The p value was adjusted by the Benjamin-Hochberg procedure for FDR <0.05.

Disease-related and drug target analysis

Lists of cancer-related genes, disease-related genes, candidate cardiovascular disease genes, and potential drug targets were obtained from the HPA database (version: 19.3). Lists of the known drug targets and the drug (therapeutic) categories of known drug-target pairs were extracted from the DrugBank database (version 5.1). The disease-related gene set, drug target gene set, and drug (therapeutic) categories enrichment analyses on the DE genes or proteins were respectively performed with a right-tailed Fisher's exact test. The p value was adjusted by the Benjamin-Hochberg procedure for FDR <0.05.

Drug repositioning and potential drug targets

To screen drug-repositioning targets and potential drug targets, four common criteria and three non-common criteria were applied. For each candidate protein, the four common criteria were as follows: (1) the average expression of the candidate protein is upregulated (log2 T/N > 0) at both G1 and G2. (2) The candidate protein is not significantly different (ANOVA, FDR >0.1) with respect to the cancer stage. (3) The candidate protein is not annotated as a lung-cancer-favorable prognostic gene in the HPA database. (4) The functions of the candidate protein in its corresponding drug are involved in cancer progressions such as proliferation, apoptosis, cell death, and cell growth. The above information for checking candidate proteins was collected via a literature survey and the Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, Redwood City, CA). The following three non-common criteria were used only for screening drug-repositioning targets: (1) the candidate protein is annotated as a known drug target in the DrugBank database, and its drug-target pair is with known pharmacological action as inhibitors. (2) The candidate protein is significantly differentially expressed (ANOVA, p value < 0.0025 and FDR <0.05; Tukey HSD, adj. p < 0.03) between G1 and G2. (3) The mean difference of the candidate protein's average expression between G1 and G2 was greater than 0.2. The following non-common criteria were used only for screening potential drug targets: (1) the candidate protein is annotated as a potential drug target in the HPA database. (2) The candidate protein is differentially expressed (ANOVA, p < 0.0065 and FDR <0.08, Tukey HSD, adj. p < 0.03) between G1 and G2. (3) The candidate protein's average expression ratios of G1 and G2 differ by more than 0.4.

Immune-related gene and proteins analyses

We acquired a list of 833 immune-related genes/proteins from the UniProt database (April 22, 2020 released) using the keyword "Immunity" (KW-0391). Comparing the immune-related gene/protein list with the DE genes and proteins, respectively, we obtained 62 DE immune-related genes and 64 immune-related proteins. We used the web-based tool Morpheus (Broad Institute, https://software.broadinstitute.org/morpheus/) to perform hierarchical clustering on the DE immune-related gene and protein expression profiles, respectively, with one minus the Pearson correlation coefficient as the pairwise distance and the average as the linkage method. We selected the results of clustering into two groups after manually comparing the results of clustering into more groups to avoid generating very small group(s), even single-entry groups. Then g:Profiler (Raudvere et al., 2019) was used to perform the GO biological process enrichment analysis. The parameters for the enrichment analysis were 25 as the minimum and 150 as the maximum numbers of IDs in each category and Benjamini-Hochberg FDR <0.01 as the significance level. The PD-L1 protein expressions of additional 38 patients that possess the *EGFR* + *TP53* mutation (n = 21 out of 110 patients) or *EGFR* mutation (n = 17 out of 110 patients) were extracted from a large-scale LUAD study by CPTAC (Gillette et al., 2020) to validate our finding regarding the difference of PD-L1 expression between the G1 (EGFR + TP53 co-mutation) and G2 (EGFR mutation) subtypes. To exclude patients possibly belonging to the G3 (multiple gene mutation) subtype among the 38 patients, we first examined the number of mutated genes that belong to the top 20 mutated genes of 88 patients in our cohort and found that only seven patients, i.e., exactly the patients in G3, had at least ten mutated genes in the top 20 mutated genes. Because selected patients from the CPTAC cohort were used to validate our finding, we applied the same rule, having at least ten mutated genes in the top 20 mutated gene list obtained from our cohort, to exclude patients possibly in G3. As a result, all of the 38 patients had less than ten mutated genes in the top 20 mutated genes and were selected. In summary, the 38 patients were selected based on the following two criteria: (1) the somatic information of EGFR and TP53 and (2) having less than ten out of the top 20 mutated genes, which were used for our somatic mutation clustering, to exclude samples possibly in the multiple gene mutation subtype (G3).

QUANTIFICATION AND STATISTICAL ANALYSIS

The details of statistical analysis are described in the results and method details subsections, and figure legends. All statistical tests were carried out using stats R package. Significance symbols correspond to p value or FDR as follows: ns ≥ 0.05, *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001.