nature portfolio

https://doi.org/10.1038/s43018-021-00259-9

Supplementary information

Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms

In the format provided by the authors and unedited

Supplementary Figures related to:

Title

Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune evasion mechanisms

Authors

Janne Lehtiö[1*$], Taner Arslan[1], Ioannis Siavelis[1], Yanbo Pan[1], Fabio Socciarelli[1], Olena Berkovska[1], Husen M. Umer[1], Georgios Mermelekas[1], Mohammad Pirmoradian[1], Mats Jönsson[2], Hans Brunnström[3,4], Odd Terje Brustugun[5,6], Krishna Pinganksha Purohit[7,8], Richard Cunningham[7,8], Hassan Foroughi Asl[9], Sofi Isaksson[2], Elsa Arbajian[2], Mattias Aine[2], Anna Karlsson[2], Marija Kotevska[2,10], Carsten Gram Hansen[7,8], Vilde Drageset Haakensen[6,11], Åslaug Helland[6,11,12], David Tamborero[1], Henrik J. Johansson[1], Rui M. Branca[1], Maria Planck[2,10], Johan Staaf[2], and Lukas M. Orre[1*]

Affiliations

[1] Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden.

[2] Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden.

[3] Department of Pathology, Laboratory Medicine Region Skåne, Lund, Sweden

[4] Division of Pathology, Department of Clinical Sciences, Lund, Lund University, Lund, Sweden.

[5] Section of Oncology, Drammen Hospital, Vestre Viken Health Trust, Drammen, Norway.

[6] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway.

[7] University of Edinburgh Centre for Inflammation Research, Institute for Regeneration and Repair, Queen's Medical Research Institute, Edinburgh bioQuarter, 47 Little France Crescent, Edinburgh EH16 4TJ, UK

[8] MRC Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh bioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK.

[9] Genomic Medicine Center, Karolinska University Hospital, Stockholm, Sweden. Clinical Genomics Facility, Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden.

[10] Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund, Sweden

[11] Department of Oncology, Oslo University Hospital, Oslo, Norway.

[12] Faculty of Medicine, University of Oslo, Norway.

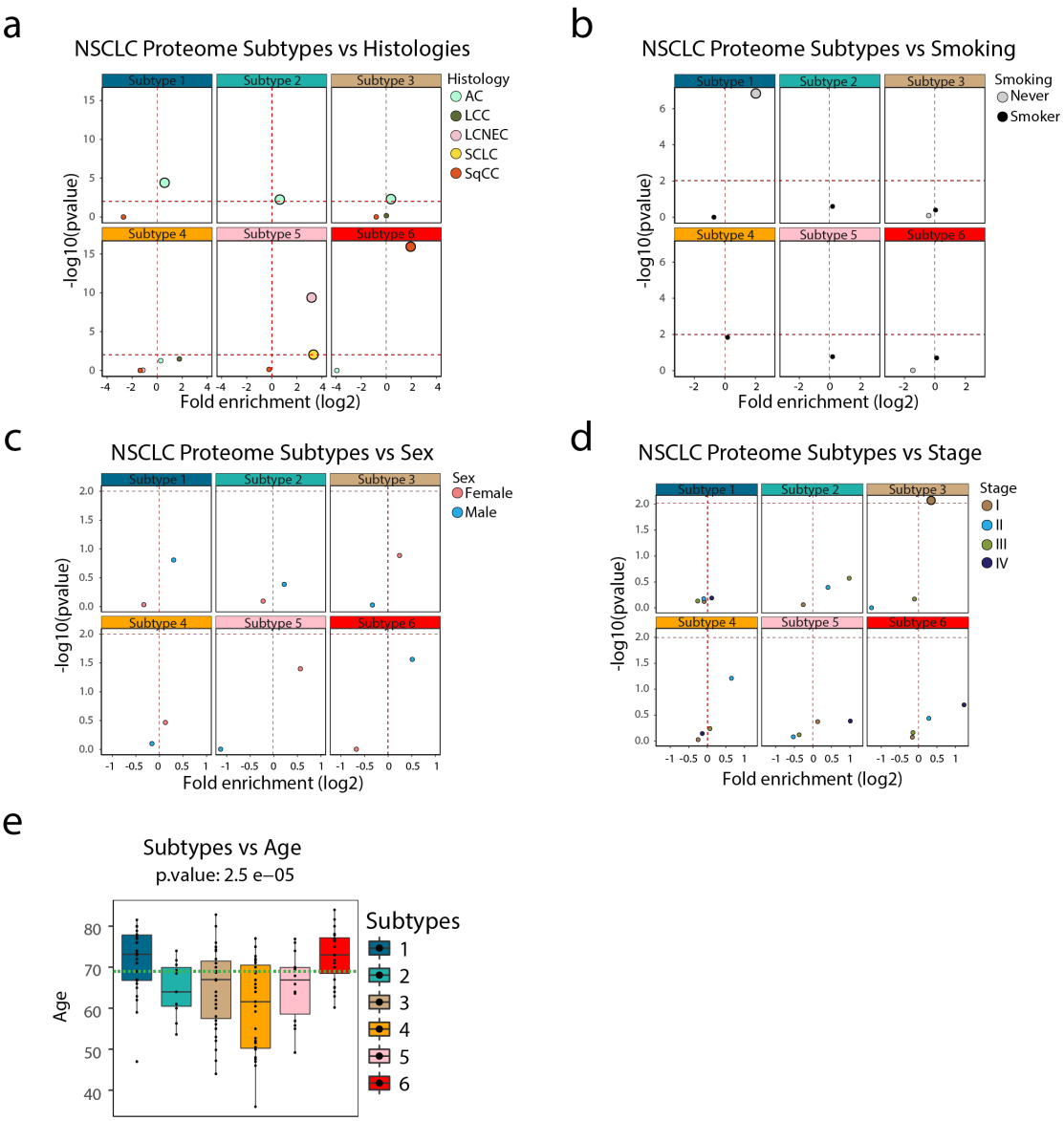
Footnotes

* These authors contributed equally.

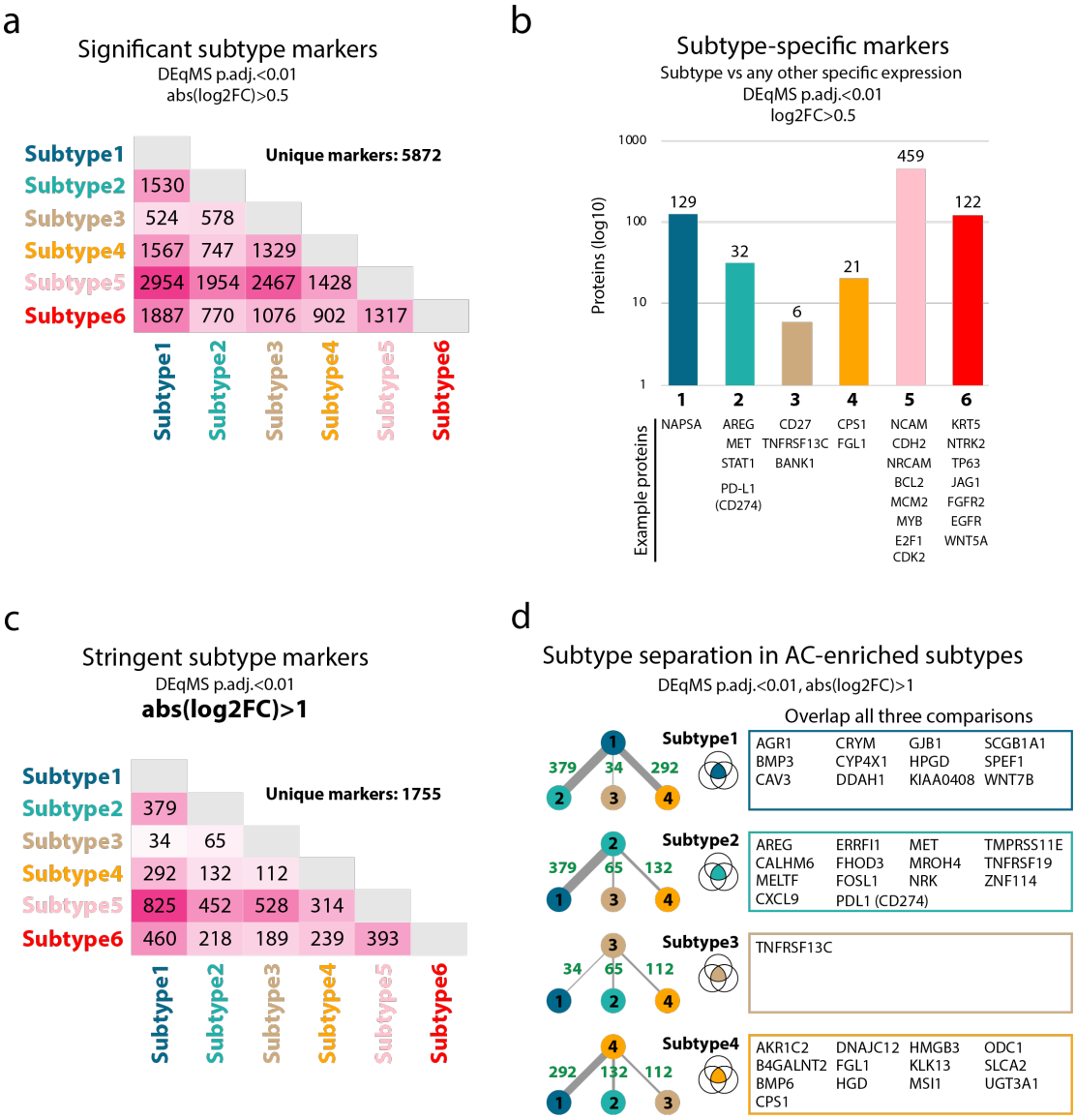
$ Corresponding author

Contact Information

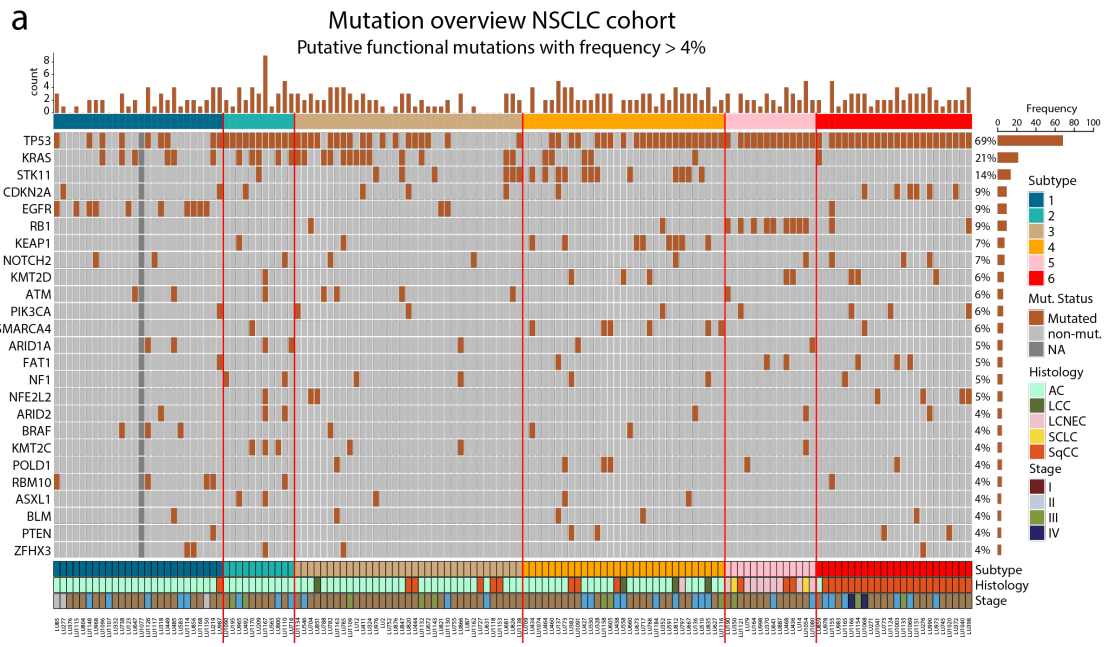
Correspondence to janne.lehtio@ki.se



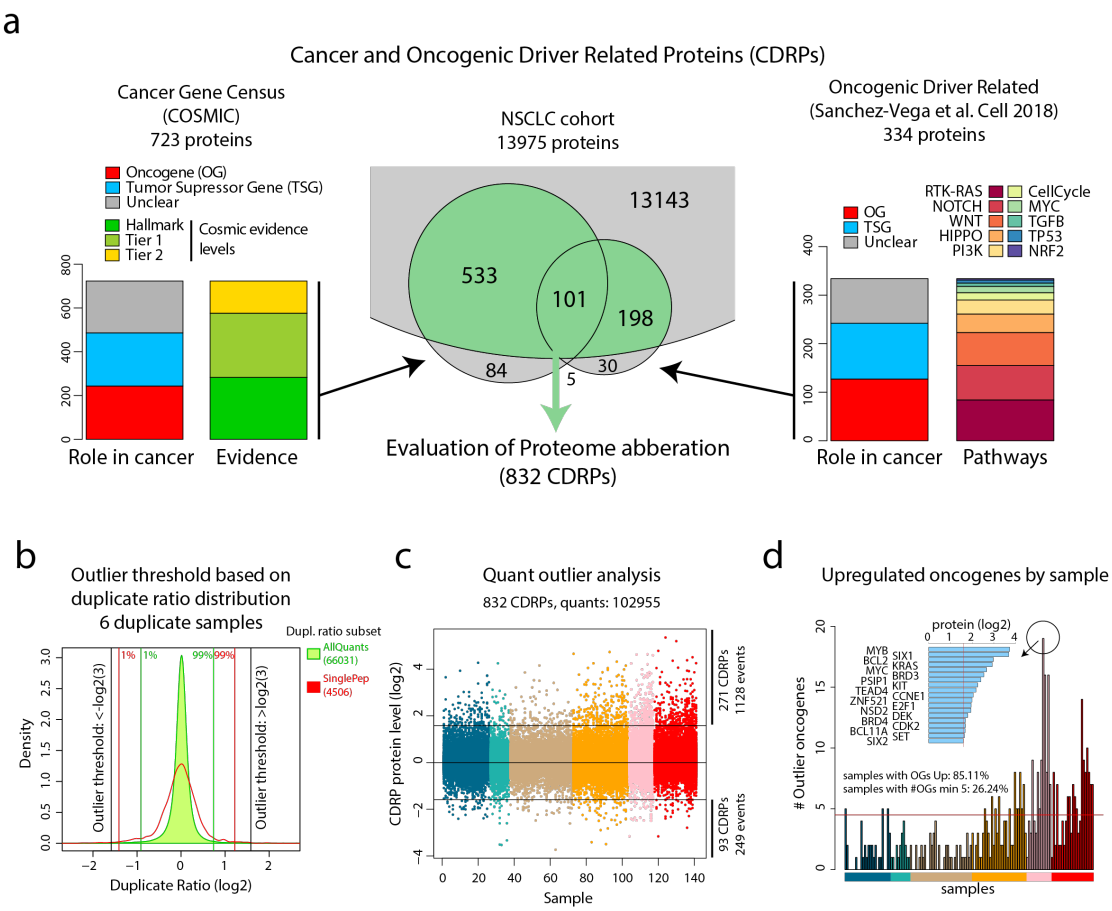
Supplementary Figure 1. Enrichments and tests for the NSCLC Proteome Subtypes. Volcano plots showing the output from enrichment tests of histology (**a**), smoking status (**b**), sex (**c**), stage (**d**). P values were calculated using one-sided hypergeometric test with Benjamini-Hochberg adjustment. Horizontal red dotted lines in all volcano plots indicate p-value=0.01. **e.** Boxplot indicating age distribution across the NSCLC Proteome Subtypes (n = 141 samples). Green dotted lines indicate cohort median. Middle line, median; box edges, 25th and 75th percentiles; whiskers, most extreme points that do not exceed ±1.5 × the interquartile range (IQR). P-value was calculated by Kruskal-Wallis test. Dunn's multiple comparison tests with Benjamini–Hochberg adjustment are available in **Supplementary Table 3.**



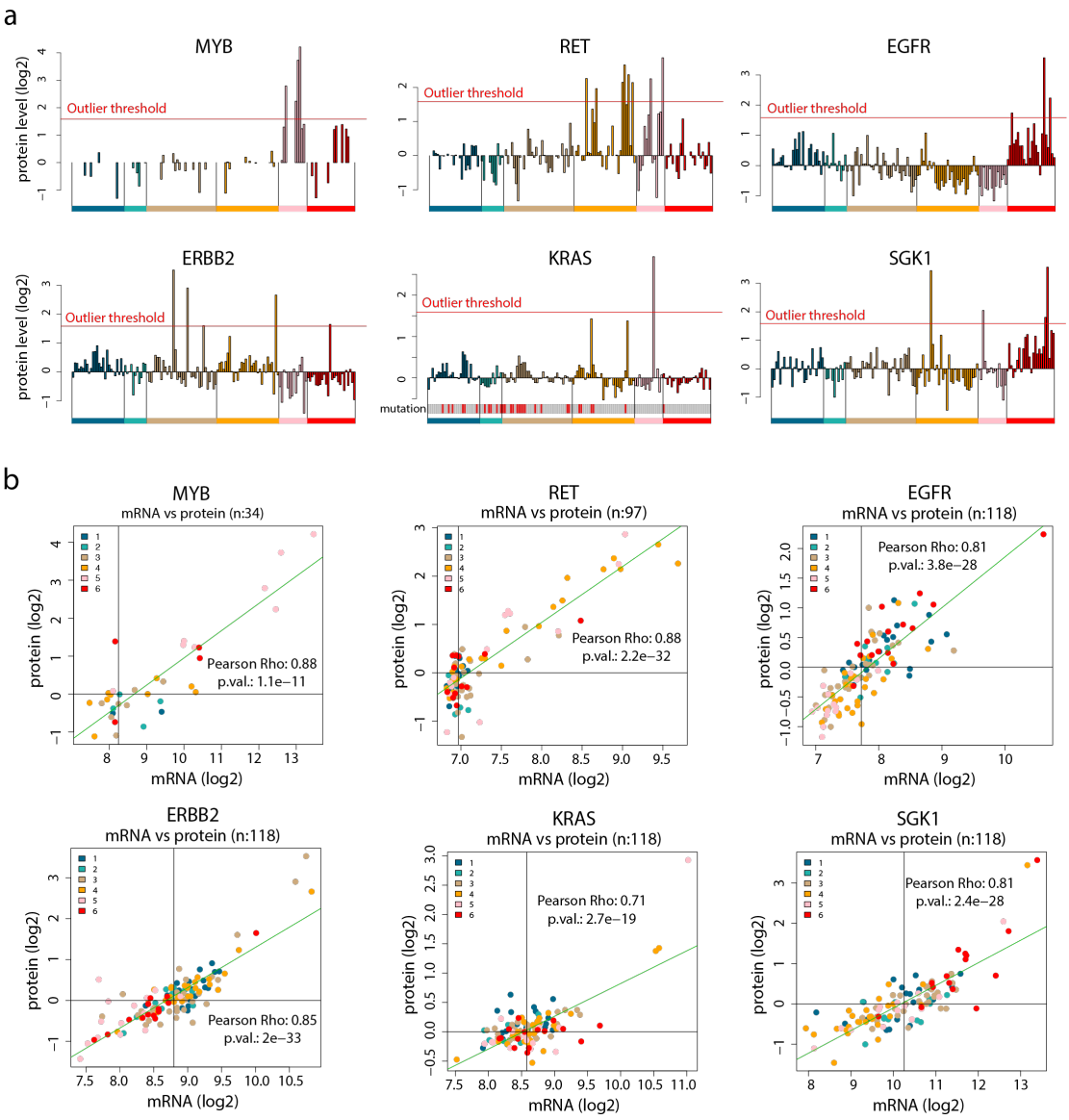
Supplementary Figure 2. NSCLC proteome subtype markers **a.** Output from DEqMS analysis to identify differentially expressed proteins between NSCLC proteome subtypes. Numbers in the plot indicate for each comparison the number of significantly different proteins (DEqMS Benjamini-Hochberg adjusted p-value<0.01, abs(log2FC)>0.5). **b.** Bar plot indicating the number of proteins with subtype specific expression (DEqMS Benjamini-Hochberg adjusted p-value<0.01, log2FC>0.5 against all other subtypes). Shown below are selected examples from each subtype. **c.** Same as in (**a**) but with a more stringent fold change cutoff (abs(log2FC)>1). **d.** Networks indicating, for each of the AC-enriched subtypes (1-4), the number of proteins significantly different to the other AC-enriched subtypes (DEqMS Benjamini-Hochberg adjusted p-value<0.01, abs(log2FC)>1). Indicated to the right are proteins that are common for all three pairwise comparisons.



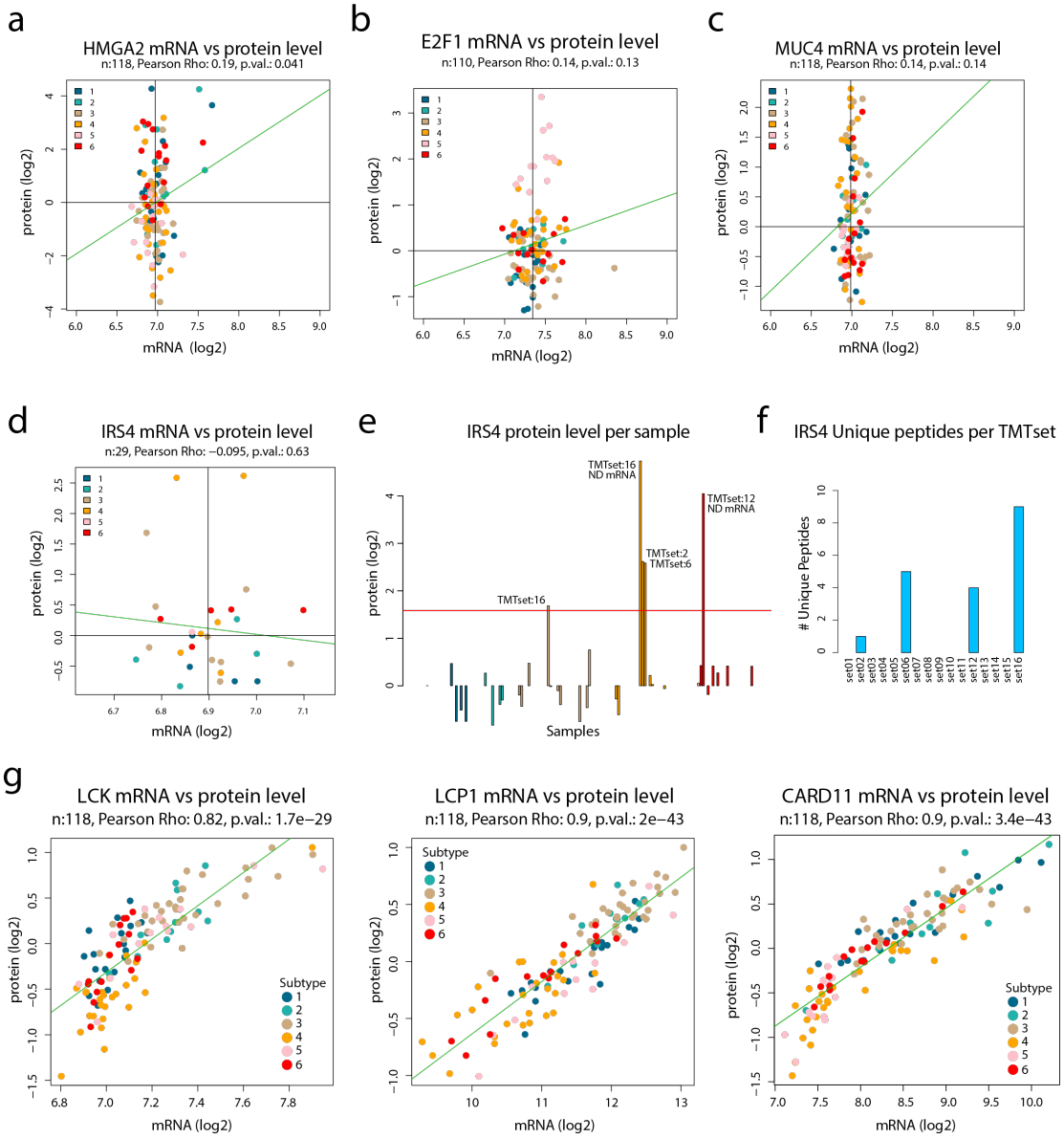
Supplementary Figure 3. NSCLC cohort panel sequencing results. **a.** Figure indicating putative functional mutations with a frequency above 4% in the 140 sequenced cohort samples, ordered by proteome subtype as in the original hierarchical clustering. The histogram on top indicates the number of common mutations per sample. The histogram on the right indicates the frequency of each mutation in the cohort. Indicated below are also cancer histology and stage.



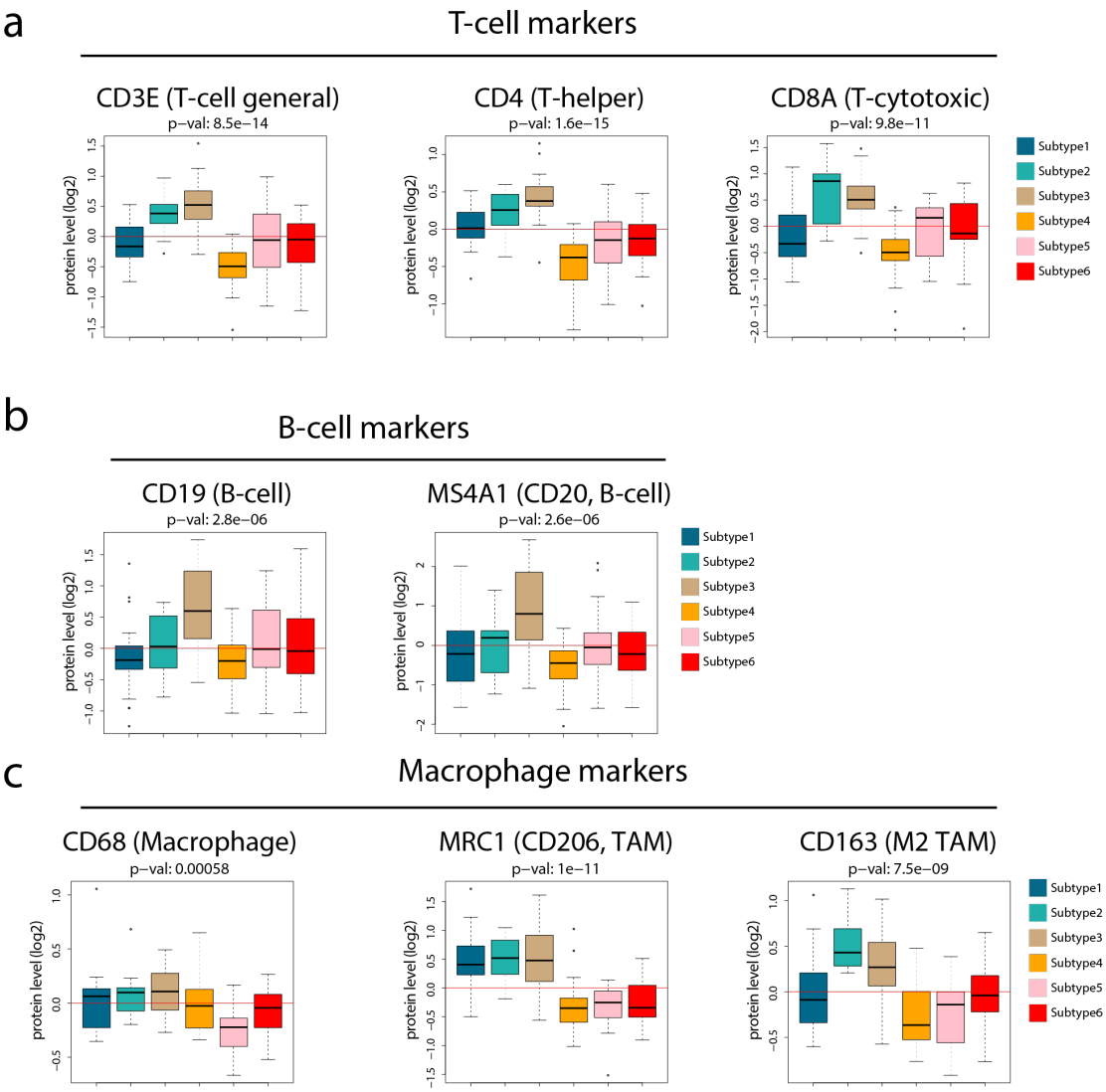
Supplementary Figure 4. CDRP quantitative outliers in NSCLC. **a.** Overview of CDRP definition and overlap with NSCLC proteome data. **b.** Density plot indicating duplicate ratios of proteins for six cohort samples that were analyzed as technical duplicates. The green line indicates all quantifications (66,031 proteins for six duplicates) and the red line indicates the subset of quantifications that were made for proteins with identification based on a single unique peptide. Vertical lines show the 1st and 99th percentile for each group. Black lines indicate the threshold for extreme expression (referred to as "outlier expression") used throughout the rest of the study. **c.** Scatterplot showing outlier expression pattern of CDRPs in the NSCLC cohort. The plot is based on 102,955 quantifications made for 832 CDRPs in the 141 cohort samples. **d.** Bar plot showing the number of overexpressed oncogenes per sample. Inset shows the protein levels of 19 oncogenes with outlier expression in a specific subtype 5 sample.



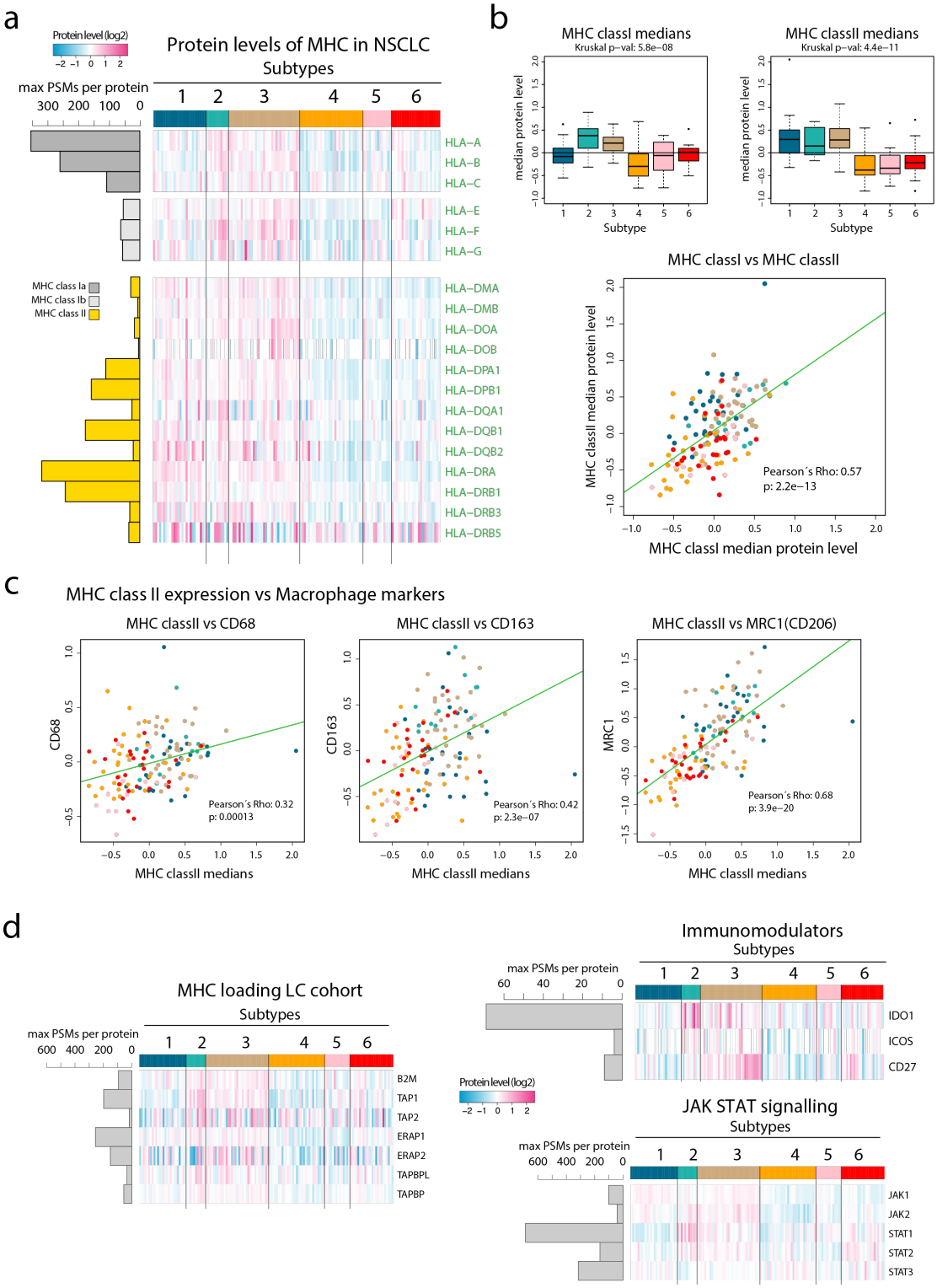
Supplementary Figure 5. Examples of oncogene outliers in the NSCLC cohort. **a.** Bar plots indicating protein patterns for the oncogenes MYB, RET, EGFR, ERBB2, KRAS, and SGK1 respectively. Indicated for KRAS is also the mutation status of KRAS. **b.** Scatter plots indicating the mRNA and protein levels of the oncogenes MYB (n = 34 samples), RET (n = 97 samples), EGFR (n = 118 samples), ERBB2 (n = 118 samples), KRAS (n = 118 samples), and SGK1 (n = 118 samples) in the NSCLC cohort. Indicated in each plot is the number of samples with quantitative information at both mRNA and protein level as well as the linear regression trendline in green. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with $n - 2$ degrees of freedom are provided.



Supplementary Figure 6. mRNA to protein correlation of CDRP outliers. **a-d.** Scatter plots indicating the mRNA and protein levels of the oncogenes HMGA2 (n = 118 samples), E2F1 (n = 110 samples), MUC4 (n = 118 samples), and IRS4 (n = 29 samples) in the NSCLC cohort. **e.** Bar plot indicating protein level of IRS4 in the cohort with indicated corresponding MS-TMT set number for samples with outlier levels. **f.** Number of unique IRS4 peptides identified per TMT set in the MS-analysis of the NSCLC cohort. **g.** Scatter plots indicating the mRNA and protein levels of LCK, LCP1 and CARD11. Indicated in each scatter plot is the number of samples with quantitative information at both mRNA and protein level as well as the linear regression trendline in green. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with $n - 2$ degrees of freedom are provided.

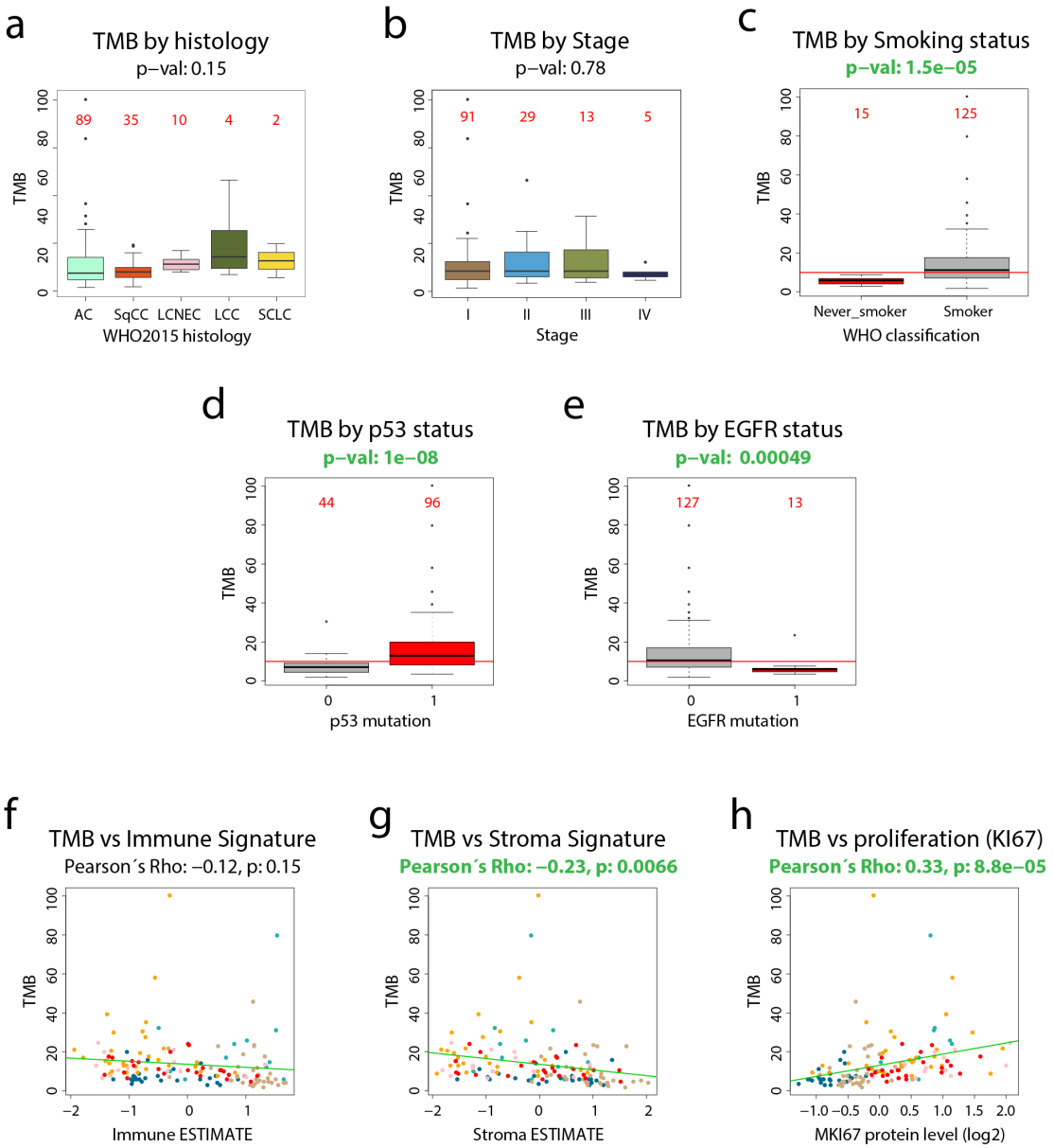


Supplementary Figure 7. Immune cell marker expression in NSCLC proteome subtypes. **a.** Boxplots indicating the protein levels of T-cell markers CD3E (n = 141 samples), CD4 (n = 141 samples), and CD8A (n = 141 samples) by proteome subtype as quantified by MS. **b.** Boxplots indicating the protein levels of B-cell markers CD19 (n = 141 samples) and CD20 (n = 141 samples) by proteome subtype as quantified by MS. **c.** Boxplots indicating the protein levels of macrophage markers CD68 (n = 141 samples), CD206 (n = 141 samples), and CD163 (n = 141 samples) by proteome subtype as quantified by MS. Middle line, median; box edges, 25th and 75th percentiles; whiskers, most extreme points that do not exceed ±1.5 × the interquartile range (IQR). P-values were calculated using two-sided Wilcoxon rank-sum test. P-values in all boxplots were calculated by Kruskal-Wallis test. Dunn's multiple comparison tests with Benjamini–Hochberg adjustment are available in **Supplementary Table 3.**

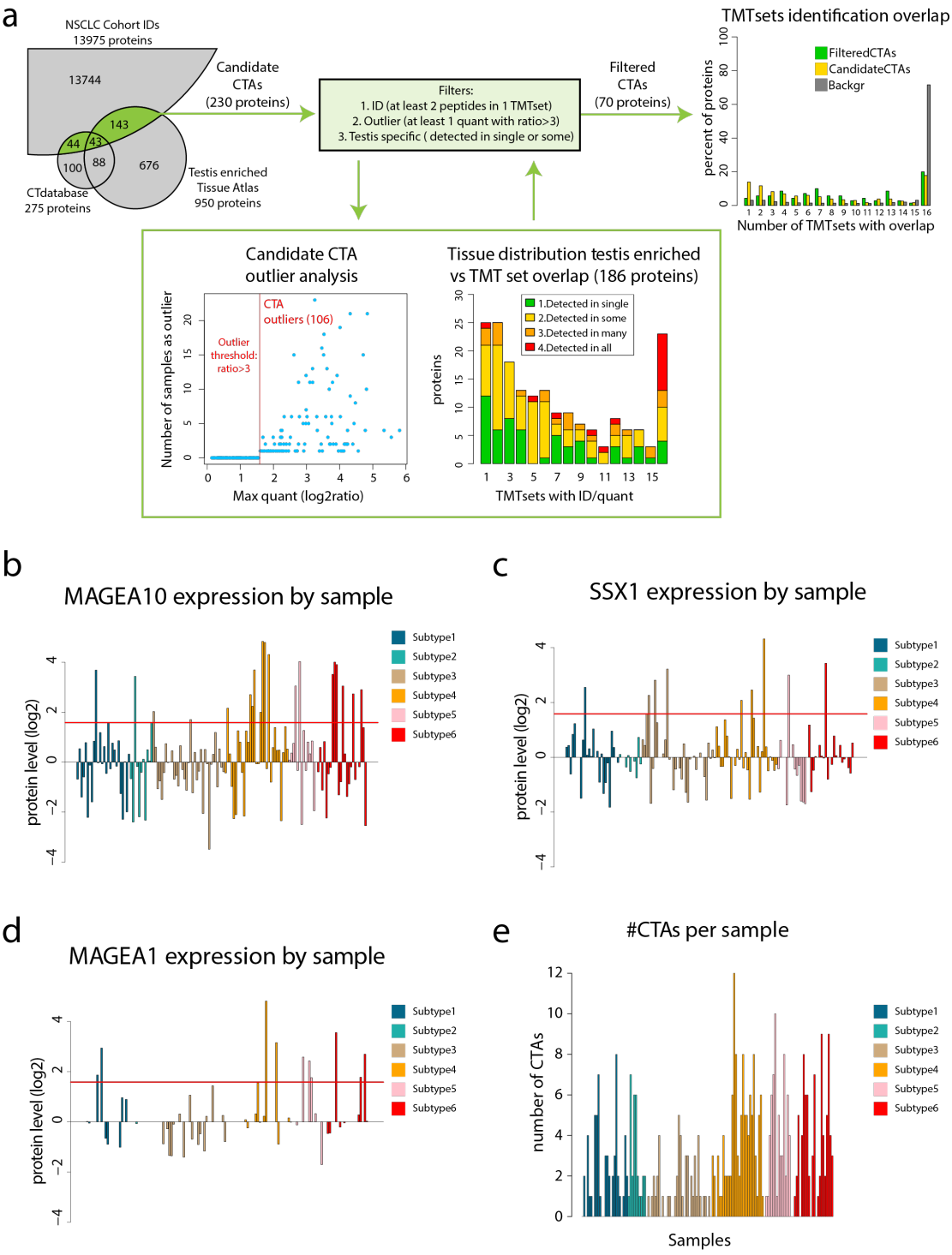


Supplementary Figure 8. Antigen processing/presentation and Immunomodulators. **a.** Heatmap indicating protein levels of HLA proteins across the NSCLC cohort samples. Indicated by bar plot on the left side is the maximum number of PSMs used for quantification of each HLA protein. **b.** Boxplots indicating the median MHC class I (left) and class II (right) protein levels by proteome subtype. Lower scatter plot indicates the median MHC class I and class II levels in each sample (n = 141 samples). **c.** Scatter plots indicating the median MHC class II protein expression plotted against the macrophage

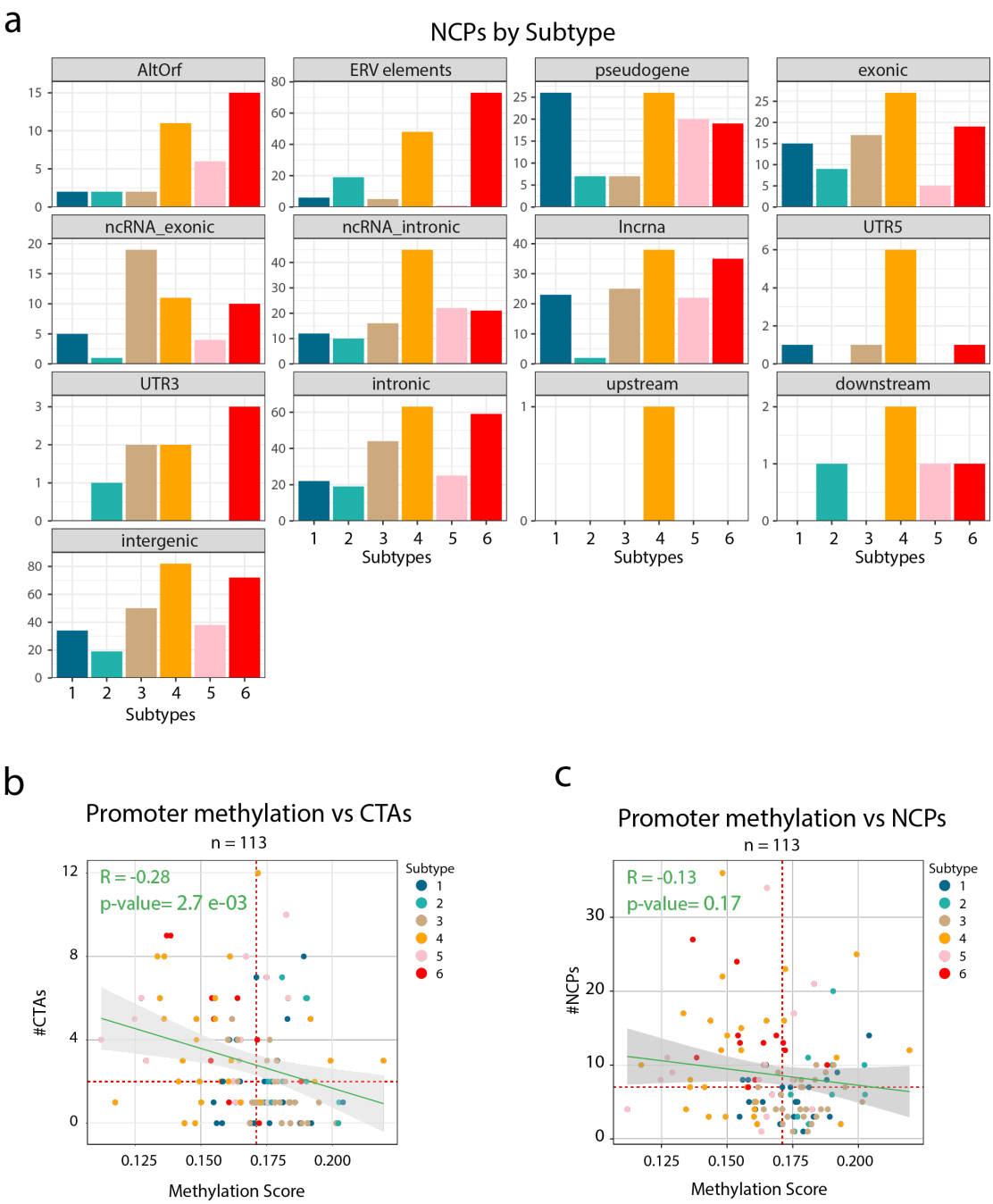
marker proteins CD68, CD163, and MRC1(CD206) (n = 141 samples). **d.** Heatmaps indicating protein levels across the NSCLC cohort samples for MHC loading proteins (left), three immune modulators (top-right) and JAK-STAT signaling pathway proteins (bottom-right). Indicated by bar plot on the left side is the maximum number of PSMs used for quantification of each protein. P-values in all boxplots were calculated by Kruskal-Wallis test. Dunn's multiple comparison tests with Benjamini–Hochberg adjustment are available in **Supplementary Table 3.** For scatter plots samples are colored by proteome subtype and a linear regression trendline is displayed in green. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with $n - 2$ degrees of freedom are provided.



Supplementary Figure 9. Tumor mutation burden (TMB) analysis in NSCLC cohort. Boxplots indicating TMB by histology (**a**) (n = 140 samples), tumor stage (n = 138 samples) (**b**), smoking status (n = 140 samples) (**c**), p53 mutations (n = 140 samples) (**d**) and EGFR mutations (n = 140 samples) (**e**). Number of samples per group is indicated in red. Middle line, median; box edges, 25th and 75th percentiles; whiskers, most extreme points that do not exceed ±1.5 × the interquartile range (IQR). P-values were calculated using Kruskal-Wallis test (**a-b**) or two-sided Wilcoxon rank-sum test (**d-e**). Dunn's multiple comparison tests with Benjamini–Hochberg adjustment are available in **Supplementary Table 3.** Scatter plots indicating TMB vs immune signature (**f**), stroma signature (**g**), proliferation (KI-67 by MS (n = 140 samples), **h**). In each plot samples are colored by proteome subtype and a linear regression trendline is displayed in green. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with $n - 2$ degrees of freedom are provided.

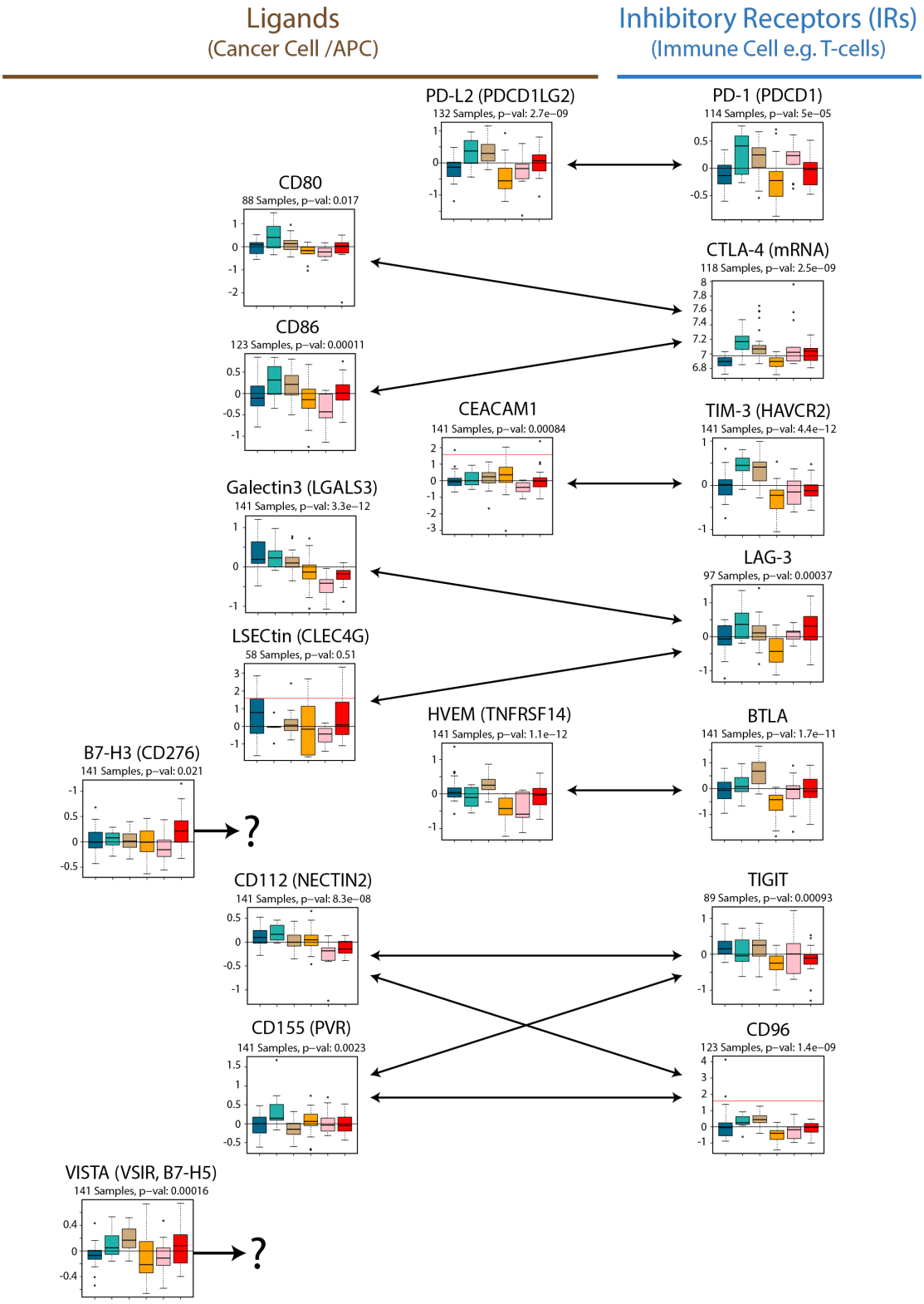


Supplementary Figure 10. Cancer-Testis antigen (CTA) expression analysis in NSCLC proteome subtypes. **a.** Overview of the CTA analysis. Candidate CTA IDs were retrieved from the CT database or the Tissue Atlas, where 230 CTAs were identified at the protein level in the NSCLC cohort. Filtering was then applied based on: at least two unique peptides per protein; outlier expression pattern; and Tissue Atlas annotation as expressed in single or some tissues. The remaining 70 CTAs used in the continued analysis showed overall low identification overlap across the NSCLC cohort as well as highly variable protein expression (examples in **b-d**), indicating sample specific, non-general protein expression of CTAs as expected. **e.** Bar plot indicating the number of CTA outliers per sample.



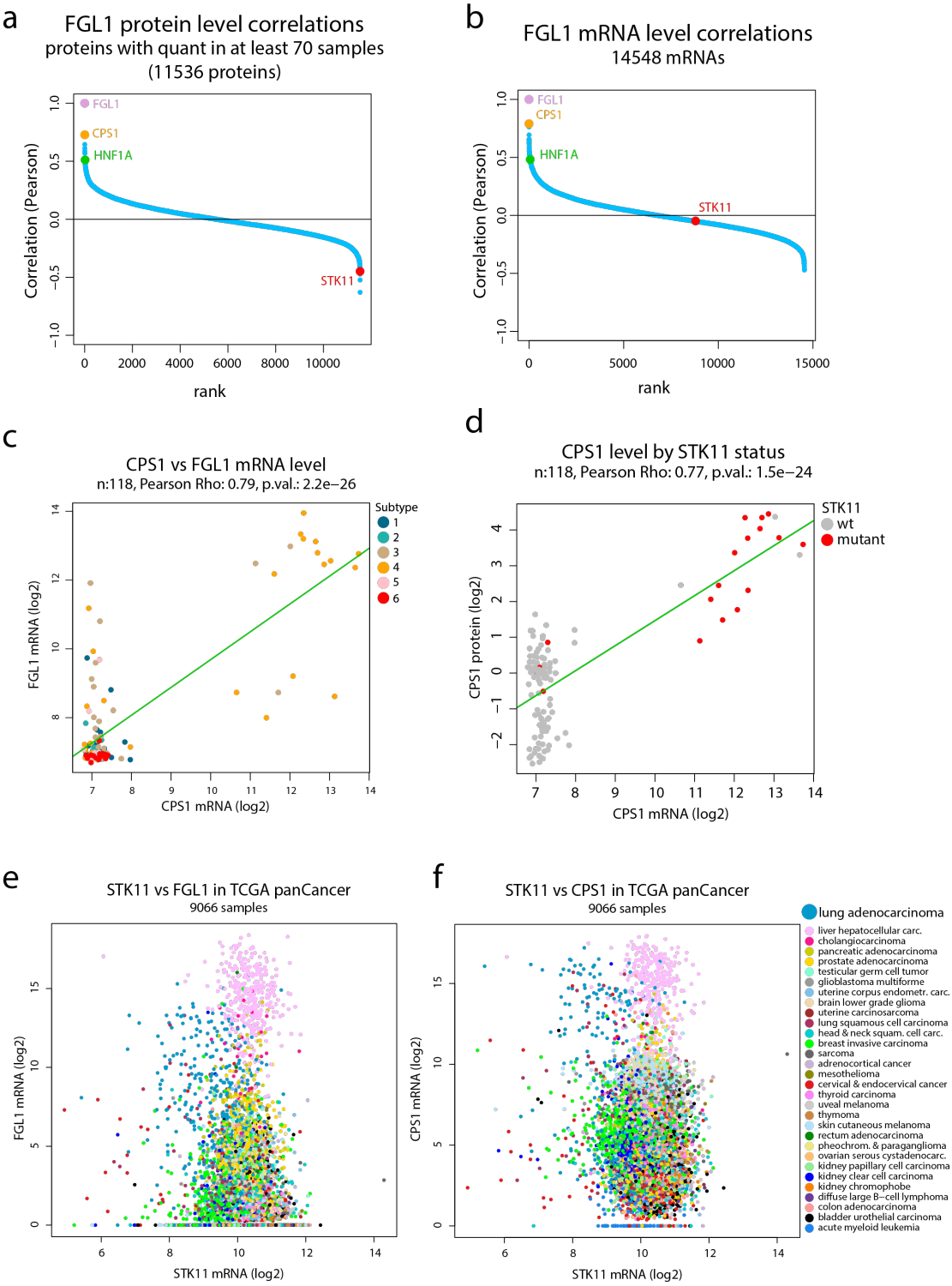
Supplementary Figure 11. Proteogenomic analysis for detection of non-canonical peptides (NCPs) in the NSCLC cohort, and CTAs and NCPs vs promoter methylation. **a.** Bar plots indicating the number of identified NCPs per subtype by mapping region type. **b.** Scatter plot indicating promoter methylation plotted against the number of CTAs per sample (n = 113 samples). **c.** Scatter plot indicating promoter methylation plotted against the number of NCPs per sample (n = 113 samples). In each plot samples are colored by proteome subtype and a linear regression trendline is displayed in green. 95% confidence intervals are shown in grey. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with *n* − 2 degrees of freedom are provided.

Immune Checkpoints in NSCLC proteome subtypes



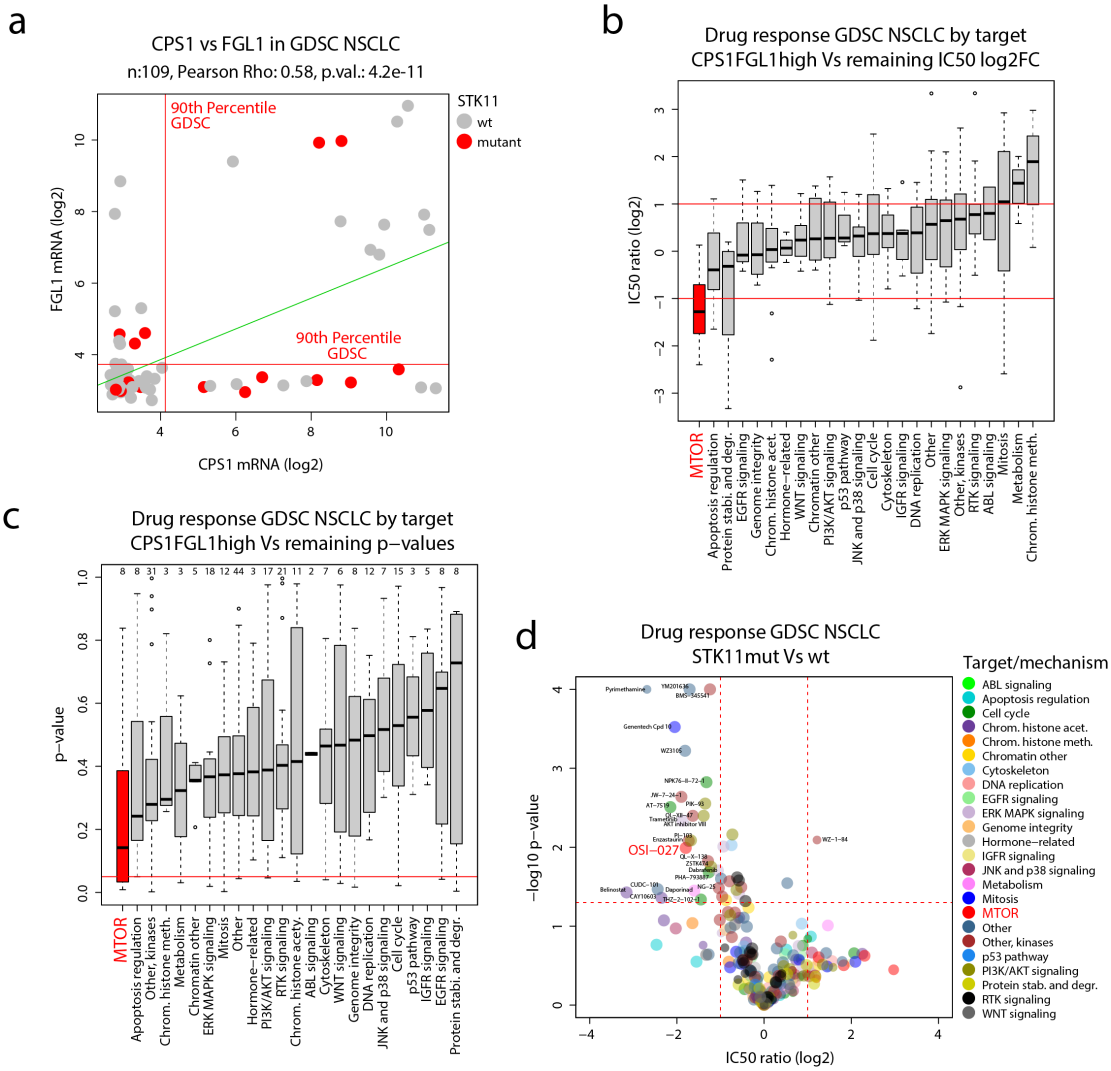
Supplementary Figure 12. Immune Checkpoints in NSCLC proteome subtypes. Boxplots indicating protein levels of inhibitory receptors (IRs) and their ligands. All values represent protein-level quantifications (log2) except for CTLA4 where mRNA levels (log2) are displayed since it was not detected by the MS data. Red lines in boxplots, where present, indicate outlier expression threshold.

Arrows indicate ligand receptor specificity according to previous literature (Andrews et al. 2019, Qin et al. 2019). Question marks indicate unknown receptors. Indicated in each scatter plot is the number of samples with quantitative information. Middle line, median; box edges, 25th and 75th percentiles; whiskers, most extreme points that do not exceed ±1.5 × the interquartile range (IQR). P-values were calculated using Kruskal-Wallis test. Dunn's multiple comparison tests with Benjamini–Hochberg adjustment are available in **Supplementary Table 3.**



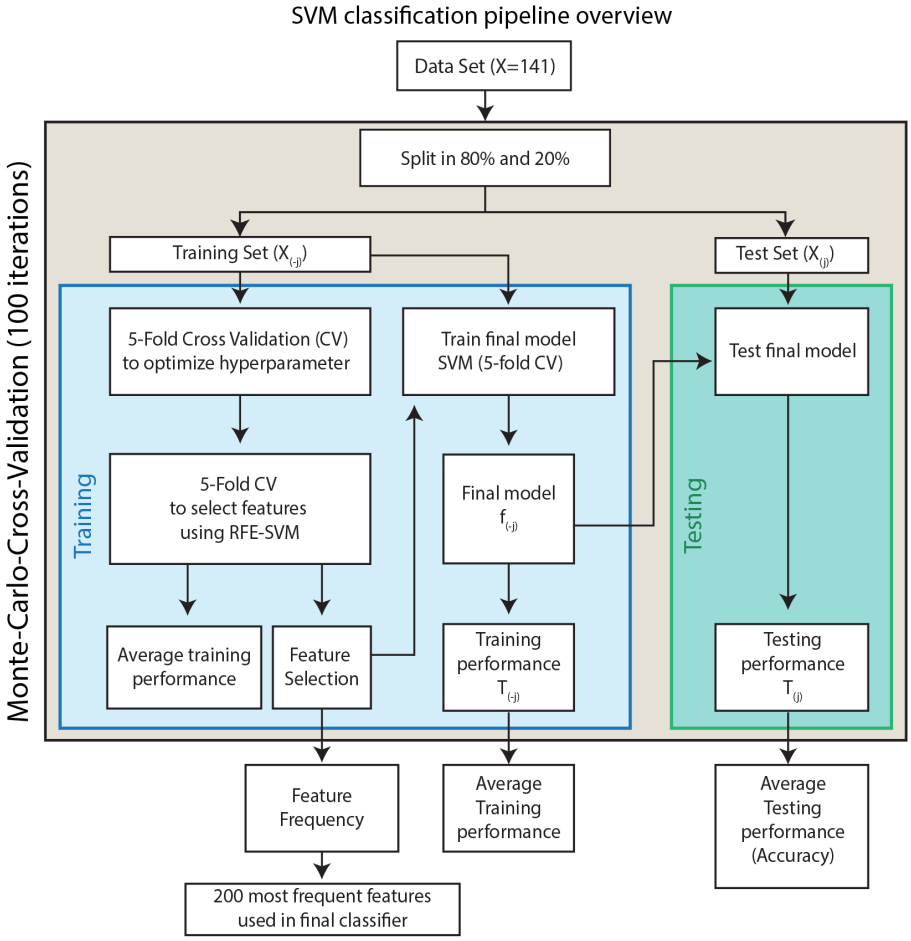
Supplementary Figure 13. FGL1, STK11 and CPS1 in NSCLC proteome landscape and TCGA pan-Cancer. **a.** Scatter plot showing ranked protein level Pearson's correlations in the NSCLC cohort. The plot includes 11,536 proteins where quantitative data was available for at least 70 samples. **b.** Scatterplot showing ranked mRNA level Pearson's correlations in the NSCLC cohort for 14548 mRNAs. **c.** Scatterplot showing CPS1 vs FGL1 mRNA levels in NSCLC cohort colored by proteome subtype (n = 118 samples). **d.** Scatterplot showing CPS1 mRNA levels vs protein levels in NSCLC cohort colored by STK11 mutation status (n = 118 samples). **e.** Scatterplot showing STK11 vs FGL1 mRNA levels in the TCGA PanCancer

dataset colored by cancer type. **f.** Scatterplot showing STK11 vs CPS1 mRNA levels in the TCGA PanCancer dataset colored by cancer type. For scatter plots **c** and **d**, a linear regression trendline is displayed in green. The associated Pearson's correlation coefficients (Rho) and two-sided p-values from *t*-distribution with $n - 2$ degrees of freedom are provided.

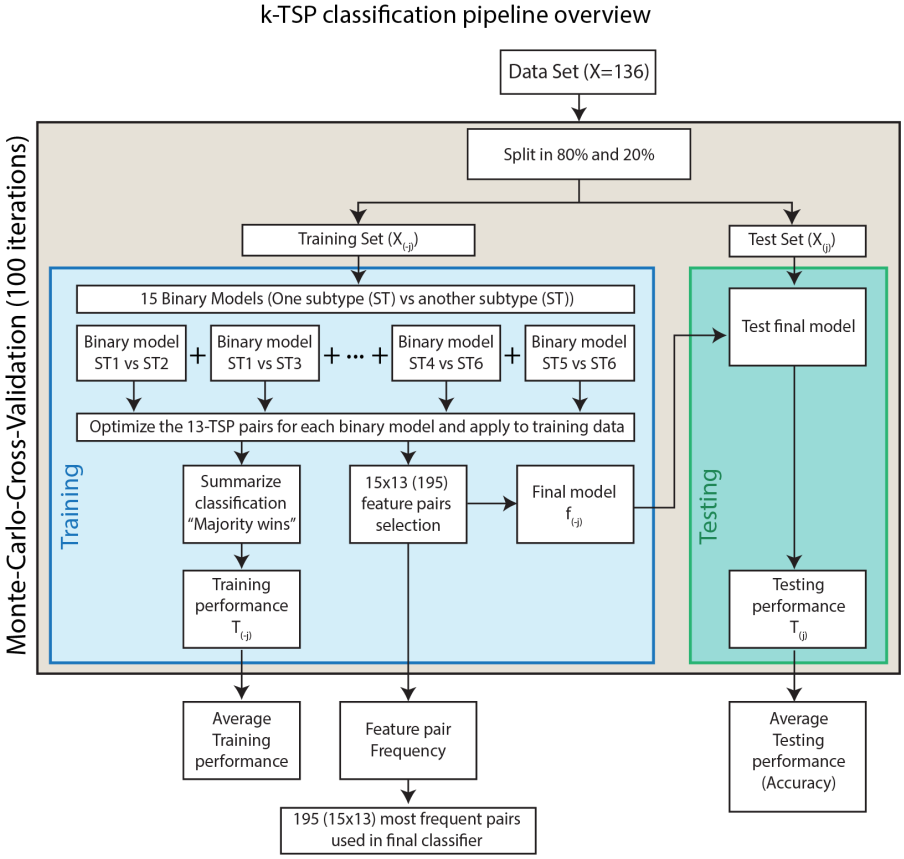


Supplementary Figure 14. FGL1 and CPS1 vs drug response in the Genomics of Drug Sensitivity in Cancer (GDSC) resource. **a.** Scatter plot showing CPS1 vs FGL1 mRNA levels in GDSC NSCLC cell lines colored by STK11 mutation status (n = 109 samples). Red lines indicate 90th percentiles of CPS1 and FGL1 mRNA expression. Linear regression trendline is displayed in green. The associated Pearson's correlation coefficient (Rho) and two-sided p-value from *t*-distribution with $n - 2$ degrees of freedom are provided. **b.** Boxplot showing the output from a differential drug response analysis between FGL1/CPS1 high NSCLC cell lines and remaining cell lines. Y-axis indicates the IC50 log2 FC by drug target group. **c.** Boxplot showing the output from a differential drug response analysis between FGL1/CPS1 high NSCLC cell lines and remaining cell lines. Y-axis indicates the p-value by drug target group as calculated by two-sided Welch's t-test uncorrected for multiple testing. **d.** Volcano plot showing the output from a differential drug response analysis between STK11 mutated NSCLC cell lines and STK11 wild-type cell lines. Y-axis indicates the -log10 p-value as calculated by Welch's t-test uncorrected for multiple testing, and x-axis indicates the IC50 log2 FC. For boxplots: Middle line, median; box edges, 25th and 75th percentiles; whiskers, most extreme points that do not exceed ±1.5 × the interquartile range (IQR).

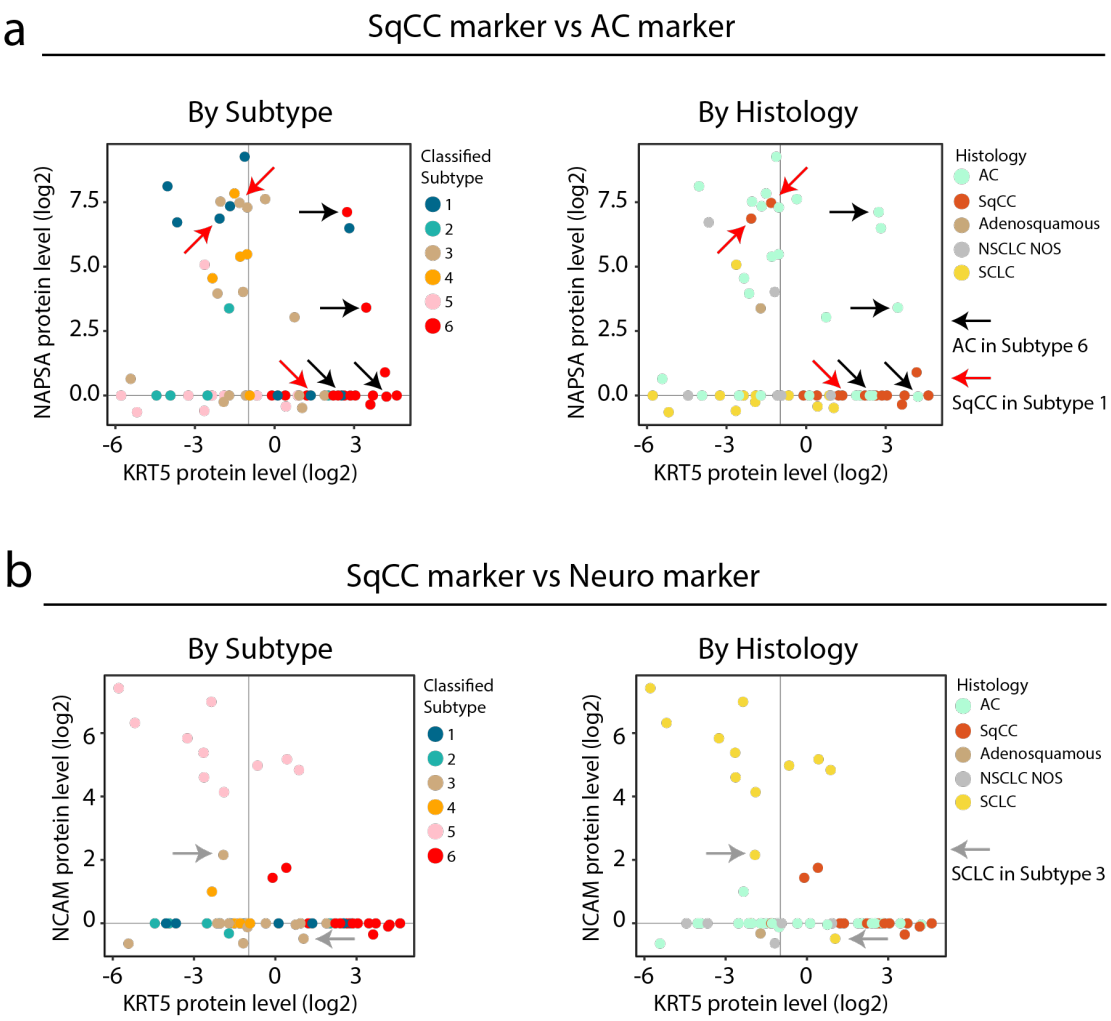
a


SVM classification pipeline overview
Data Set (X=141)
Split in 80% and 20%
Training Set (X(-j))
Test Set (X(j))
Monte-Carlo-Cross-Validation (100 iterations)
Training
5-Fold Cross Validation (CV) to optimize hyperparameter
Train final model SVM (5-fold CV)
5-Fold CV to select features using RFE-SVM
Final model f(-j)
Average training performance
Feature Selection
Training performance T(-j)
Testing
Test final model
Testing performance T(j)
Feature Frequency
Average Training performance
Average Testing performance (Accuracy)
200 most frequent features used in final classifier

b


k-TSP classification pipeline overview
Data Set (X=136)
Split in 80% and 20%
Training Set (X(-j))
Test Set (X(j))
Monte-Carlo-Cross-Validation (100 iterations)
Training
15 Binary Models (One subtype (ST) vs another subtype (ST))
Binary model ST1 vs ST2
Binary model ST1 vs ST3
Binary model ST4 vs ST6
Binary model ST5 vs ST6
Optimize the 13-TSP pairs for each binary model and apply to training data
Summarize classification "Majority wins"
15x13 (195) feature pairs selection
Final model f(-j)
Training performance T(-j)
Testing
Test final model
Testing performance T(j)
Average Training performance
Feature pair Frequency
Average Testing performance (Accuracy)
195 (15x13) most frequent pairs used in final classifier

Supplementary Figure 15. Support-vector machine (SVM)-based cohort classifier for NSCLC subtype classification. **a.** Overview of the SVM classification optimization. Briefly, Monte Carlo cross-validation (MCCV) with 100 iterations was used to estimate classifier accuracy, select the most important features and to build the final classifier. For each iteration the cohort samples were split into a training set and a testing set, and RFE-SVM was applied to select the 200 most important features. After training, the accuracy of the classifier was estimated using the test set samples. The overall accuracy was reported as the average accuracy of the 100 iterations, and the most frequently used 200 features were used to build the final model. **b.** Overview of the k-TSP classification optimization. Briefly, Monte Carlo cross-validation (MCCV) with 100 iterations was used to estimate classifier accuracy, select the most important feature pairs and to build the final classifier. For each iteration the cohort samples were split into a training set and a testing set, and 15 binary models were built, each based on 13 feature pairs (k), selected based on accuracy in test data.



Supplementary Figure 16. k-TSP based classification vs Histology in late-stage NSCLC cohort. **a.** Scatterplot indicating Napsin A (AC marker) vs Keratin 5 (SqCC marker) levels in the classified subset of the late-stage NSCLC cohort as quantified by DIA-MS. Left plot is color-coded by classified subtype and right plot by histology. Indicated by arrows in the plots are cases with unexpected classification output. **b.** Same as in a. but for Keratin 5 (SqCC marker) vs NCAM1 (Neuronal lineage marker). Indicated by arrows in the plots are cases with unexpected classification output.