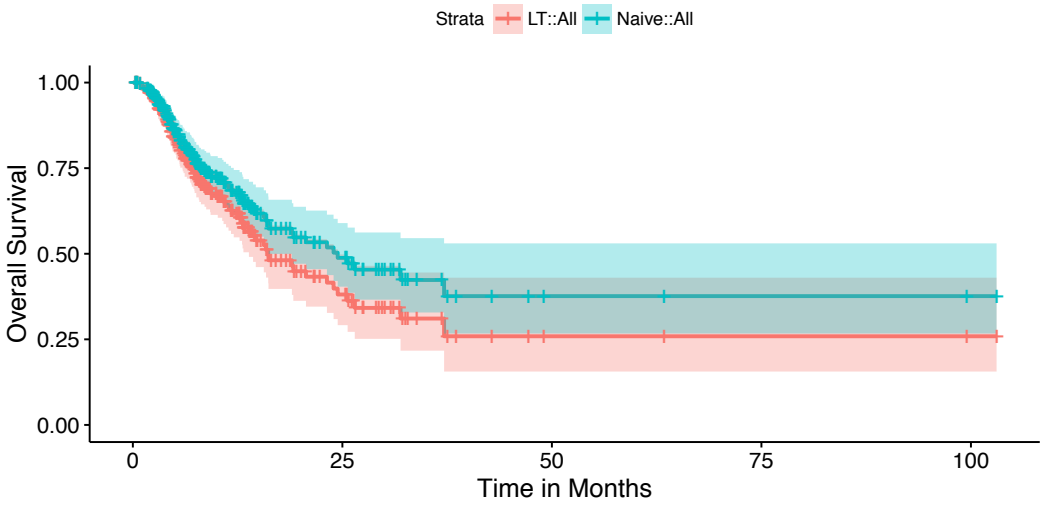
Supplementary Methods

Patients. Consecutive patients with metastatic or recurrent lung adenocarcinomas for which MSK-IMPACT data were available were included. Electronic medical record was used to identify patient clinical factors as well as survival outcomes. Data collection was approved by the MSKCC Institutional Review Board/Privacy Board. Overall survival (OS) was defined as the time from date of diagnosis of advanced disease (stage IV or recurrent cancer) until date of death or last follow-up. In this data set, the majority (68%) of the tumors in our cohort were biopsied and sequenced within 30 days of diagnosis of metastatic disease. However, a fraction (21%) of the tumors were sampled and sequenced more than six months from the met recurrence date, with 16% more than a year and 8% more than 2 years. We adjusted the late entry by left-truncation as described in the next section.

"Late entry" refers to these patients with older samples used for sequencing analysis at a later time in a minority of patients. Those patients with older samples taken at initial diagnosis of advanced lung cancer were "immortal" from their initial sampling time to the time of referral for MSK-IMPACT sequencing. This interval can be long for a small fraction of patients (8% with a delayed interval more than 2 years as mentioned in Supplementary Material), introducing survival bias. We show that standard survival analysis without adjusting for the bias can lead to over-optimistic survival estimates. For example, the standard Kaplan-Meier survival estimate without adjusting for the survival bias (termed "Naïve" analysis) for KRAS-mutant patients was biased upward (blue curve in Figure shown below) compared to that after adjusting for the survival bias (red curve). The median survival estimated from unadjusted analysis was 24 months for KRAS-mutant advanced lung cancer, significantly higher from

previously reported in the literature[1,2]. We observed similar upward bias for EGFR-mutant advanced lung cancer as well.



	MedianOS	95%CI	3Ysurvival	5Ysurvival
Left-truncated KRAS = 1	16	(0.41,0.59)	0.31	0.26
Naive KRAS = 1	24	(0.42,0.6)	0.42	0.38
Left-truncated KRAS = 0	25	(0.45,0.56)	0.39	0.24
Naive KRAS = 0	38	(0.45,0.56)	0.54	0.42

We used the left-truncation method discussed in Kalbfleisch and Prentice [3] to adjust for this bias. This entails setting up Left-truncated and Right-censored (LTRC) data to construct the Surv function in the R penalized package for Lasso-penalized Cox regression. The LTRC data included (L_i, R_i, δ_i, x_i) where L_i was time from diagnosis of advanced lung cancer to time of MSK-IMPACT sequencing; R_i was time from diagnosis to death or last follow-up; δ_i was the censoring indicator, and x_i denotes covariate vector. We show the left-truncation analysis was effective in eliminating the survival bias and the median OS of 16 months was now close to what's reported in the literature for KRAS-mutant advanced lung adenocarcinomas. We also note that although left-truncation was necessary in this analysis to adjust for survival bias, this bias will likely diminish as clinical sequencing increasingly becomes a part of routine care and adopted more widely at academic centers and community oncology setting with more tumors sampled and sequenced in a timely and efficient manner. The

discussion on left truncation analysis has been included in the Supplementary Methods section.

OncoCast. We developed OncoCast, a computational tool for integrating tumor sequencing and clinical data for survival prediction in cancer. OncoCast implements a lasso-penalized Cox regression as the core algorithm for deriving prediction rules for overall survival. In this study, the survival time was calculated from the time of metastatic disease to the time of death or last follow-up. The hazard function for overall survival at time t can be written as:

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

$$= \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is the vector of regression coefficients, and $\mathbf{x} = (x_1, ..., x_p)$ is the vector of genomic variables. In our analysis, the genomic predictors include mutation, copy number, and fusion detected from the MSK-IMPACT sequencing assay. In our data, tumor sequencing was performed using MSK-IMPACT assay including version 1 (341 genes), version 2 (410 genes), and version 3 (468 genes) in different sets of samples. Genes overlapping all platform versions and altered at least once in the sample cohort were included for this analysis.

The estimation of the parameter $\boldsymbol{\beta}$ is achieved though maximization of the partial likelihood

$$L(\beta) = \Pi_{r \in D} \frac{\exp(\beta^T x_r)}{\Sigma_{j \in R_r} \exp(\beta^T x_j)}$$

where D is the set of indicies for events (e.g., death), and R_r is the set of indices of the individuals at risk. Denote the log partial likelihood by $l(\beta) = \log L(\beta)$, Tibshirani [4] proposed penalized Cox regression through constrained likelihood optimization as

$$\hat{\beta} = argmax\ l(\beta), \text{ subject to } \left\|\beta_j\right\|_1 \leq s.$$

The l_1-norm constrained optimization shrinks the regression coefficients toward zero to achieve sparse models. Such shrinkage renders more interpretable prediction rules and reduces the chance of over-fitting. The amount of shrinkage was determined through a 5-fold cross-validation. A combination of gradient ascent and Newton-Raphson algorithm was used for optimization using the R penalized package[5]. Left-truncation was adjusted using the Kaplan-Meier estimator of the survival function Surv**(**t_1, t_2, status**)** in which t_1 is time interval between diagnosis and time of tumor biopsy/sequencing and t_2 is time to death or last follow-up.

In the OncoCast pipeline, the core penalized Cox regression is wrapped in an ensemble learning framework using a repeated sample-splitting approach. In particular, we generated m different sample splits into a training (two-thirds) and test (one-third) set from the entire cohort of sample size n. An l_1-penalized Cox regression model was fitted and tuned in each of the m training cohorts using a 5-fold cross-validation and then applied to derive a predicted risk score for the test cohort samples. We added a small l_2 penalty ($0.01 \times l_1$) for improved numerical stability in the presence of correlated features. The risk score takes the form of the linear predictor $\eta = \beta\mathbf{x}$. In a sparse model, many of the coefficient estimates will be exactly zero and thus reducing the variance of prediction and allows variable selection

We set $m = 200$ to obtain an averaged risk score for the cohort, which was then scaled between 0 and 10 for interpretability. A sensitivity analysis on m was conducted in which we varied m from 100 to 1000, which showed that the prediction performance was not significantly affected by the choice of m. In each of the $m = 200$ penalized models, we recorded the genes with non-zero coefficient estimate. These genes had non-zero weights in the linear predictor and contributed to the prediction. For each gene, we calculated the selection frequency (proportion of times with non-zero coefficient estimates) across the m models as a measure of the relative importance for individual gene alterations.

Allele-specific copy number and clonal heterogeneity analysis. Allele-specific copy number and clonal heterogeneity analysis were performed using the FACETS algorithm[6] taking into account tumor purity and ploidy. The details FACETS analysis of whole-exome and MSK-IMPACT sequencing data had been described in several previous studies[7-9] and will not be repeated here.

Mutation clonality and multiplicity. Carter et al.[10] introduced the cancer cell fraction and multiplicity concept for somatic point mutation. Cancer cell fraction (CCF) is the fraction of tumor cells carrying the mutation. CCF close to 1 indicates that the mutation is clonal in the tumor sample. Subclonal mutations have low CCF. For somatic mutations, multiplicity refers to the number of mutant copy. Denote the major and minor integer copy-number as m, n, and the possible multiplicity of the somatic mutation $s \in \{1,\ldots,m\}$. For clonal somatic mutations, the variant allele frequency can be written as

$$VAF = \frac{\rho \tau s}{s(1-\rho) + \rho t}$$

where ρ is tumor purity, τ is cancer cell fraction, *s* denotes mutant copy, and $t = m + n$ denotes total copy number at the mutant locus in the tumor. Here given $\hat{\rho}, \hat{\tau}$ obtained from FACETS, we derived an empirical estimate of *s* that minimizes $abs(\tau - 1)$. The ratio of mutant copy over the total copy number ($r = s/t$) was summarized for each mutation in each sample. For heterozygous loss regions, $r = 1$. For regions with segmental copy number gains, a high mutant copy ratio r indicates that the point mutation likely preceded the segmental duplications.

Each mutation was classified into clonal and subclonal status using the cancer cell fraction (CCF). Binomial exact confidence interval (CI) was calculated around the point estimate of CCF. Mutations with lower bound of 95% CI ≥ 75% were classified as clonal. Mutations with CCF ≥ 80% and lower bound of 95% CI below 75% were classified as likely clonal. Mutations with CCF < 80% and lower bound of 95% CI <75% were classified as subclonal.

Mutation burden and signature analysis. Mutation burden was calculated as the number of nonsynonymous mutations per sample divided by the total genomic coverage of the MSK-IMPACT platform. Mutation signatures were identified for each sample according to distribution of the six substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) and the bases immediately 5' and 3' of the mutated base, producing 96 possible mutation subtypes. Decomposition analysis was applied to map the mutation pattern in each sample to the 30 signatures that had been previously described[11]. Each signature was assigned a weight that corresponded to the percentage of mutations explained by each given signature.

Software availability. The OncoCast R package is available at https://github.com/shenmskcc/OncoCast. An interactive web interface was developed using R Shiny with two main functions: GeneView and PatientView. In GeneView, users can interactively explore gene importance and co-mutation pattern by risk groups. The relative importance of a gene is measured by the selection frequency and the average regression coefficients that quantify how much the gene weighs in the risk score and overall prediction. The two measures can be interactively visualized in the volcano plot. In PatientView, users can type in a patient's mutational profile and specify the clinical characteristics. The genomic risk score (averaged across the m penalized models) along with the predicted probability of survival at different time marks will be calculated and viewable in a dynamic plot. The Shiny app is available at https://github.com/shenmskcc/LungIMPACT.

References

1. Yu HA, Sima CS, Hellmann MD, et al: Differences in the survival of patients with recurrent versus de novo metastatic KRAS-mutant and EGFR-mutant lung adenocarcinomas. Cancer 121:2078-2082, 2015

2. Yu HA, Sima CS, Shen R, et al: Prognostic impact of KRAS mutation subtypes in 677 patients with metastatic lung adenocarcinomas. Journal of Thoracic Oncology 10:431-437, 2015

3. Kalbfleisch JD, Prentice RL: The statistical analysis of failure time data (ed 2nd). Hoboken, N.J., J. Wiley, 2002

4. Tibshirani R: The lasso method for variable selection in the Cox model. Statistics in medicine 16:385-395, 1997

5. Goeman JJ: L1 penalized estimation in the Cox proportional hazards model. Biometrical journal 52:70-84, 2010

6. Shen R, Seshan VE: FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic acids research 44:e131-e131, 2016

7. Morris LGT, Chandramohan R, West L, et al: The molecular landscape of recurrent and metastatic head and neck cancers: insights from a precision oncology sequencing platform. JAMA oncology 3:244-255, 2017

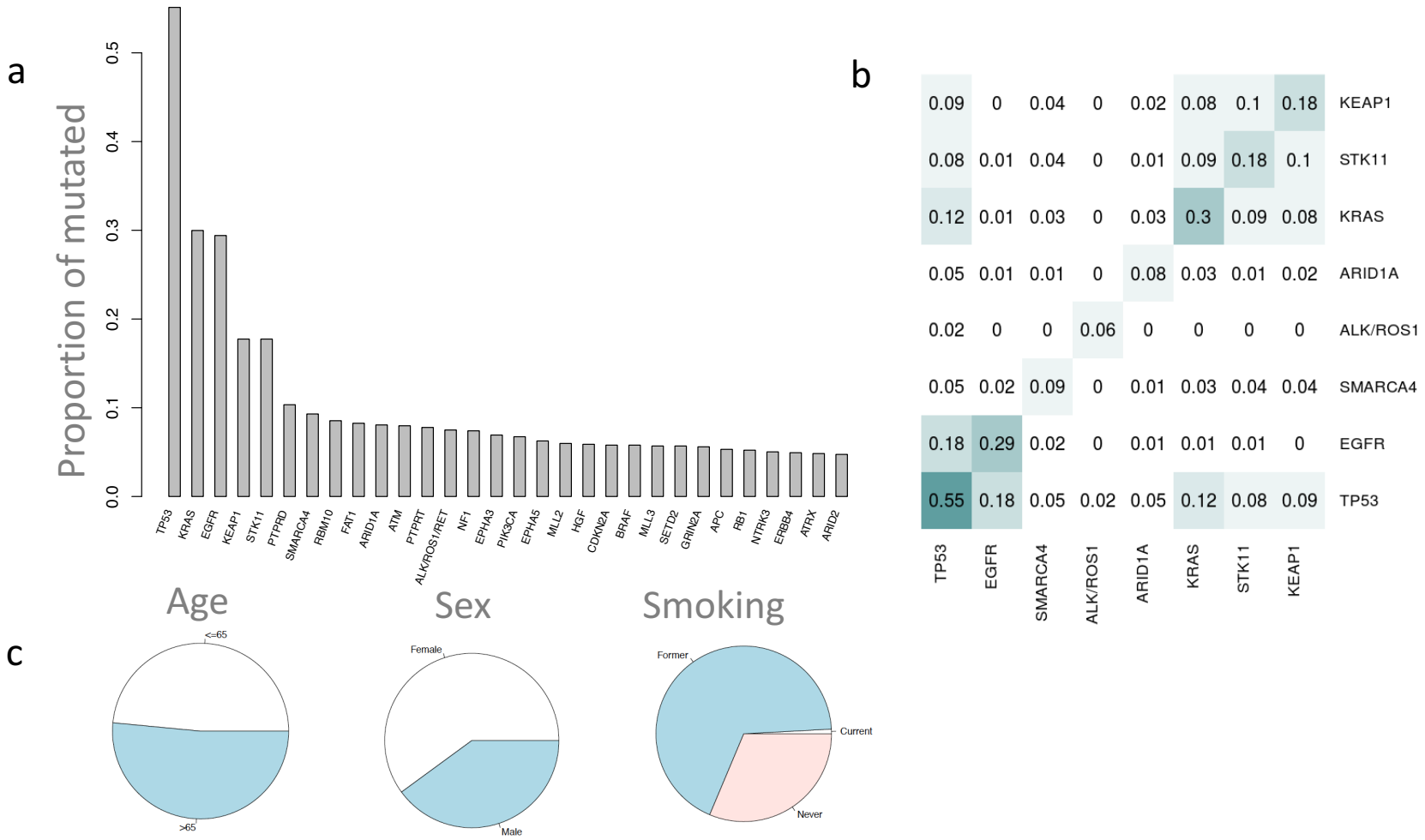
8. Paik PK, Shen R, Won H, et al: Next-generation sequencing of stage IV squamous cell lung cancers reveals an association of PI3K aberrations and evidence of clonal heterogeneity in patients with brain metastases. Cancer discovery 5:610-621, 2015

9. Riaz N, Havel JJ, Makarov V, et al: Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell 171:934-949, 2017

10. Carter SL, Cibulskis K, Helman E, et al: Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology 30:413-421, 2012

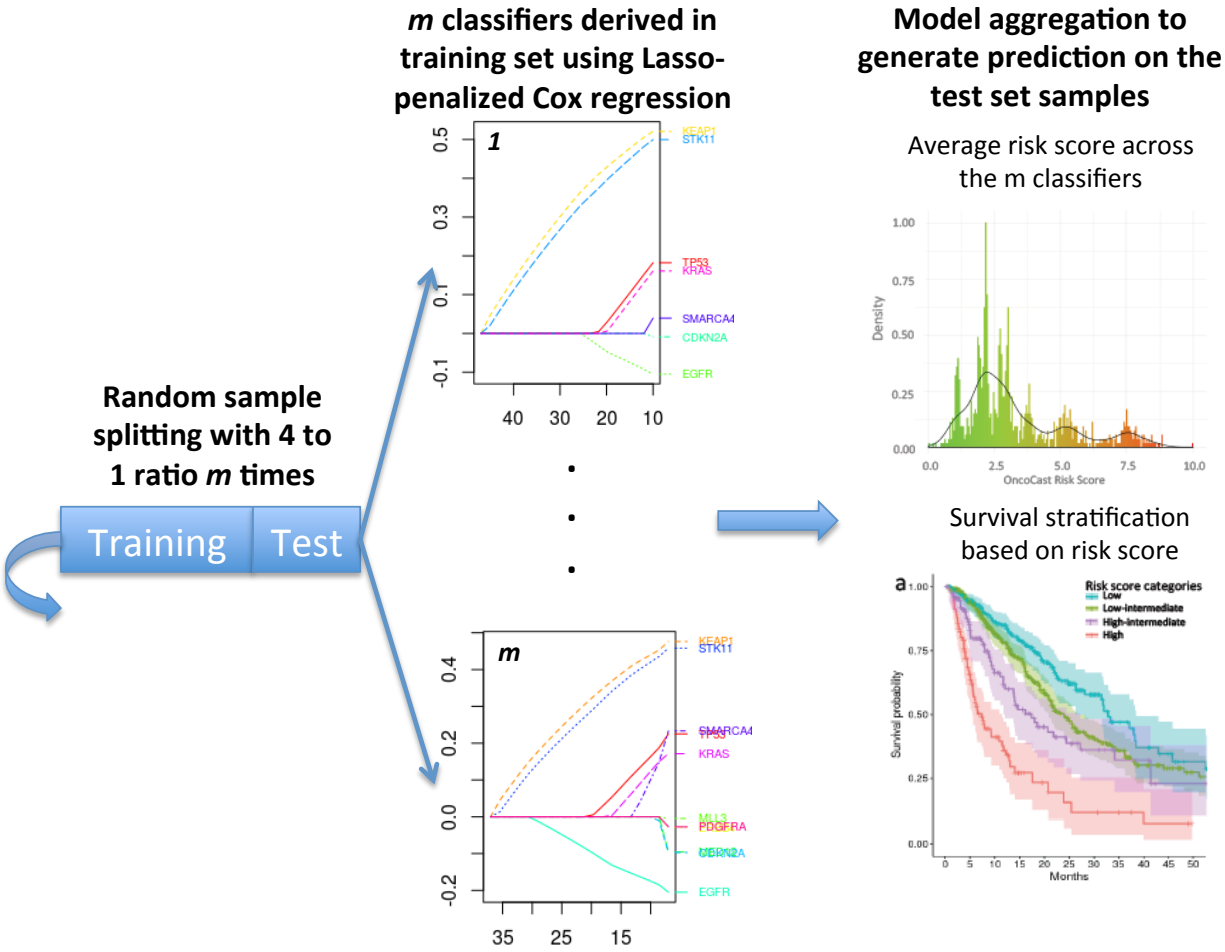
11. Alexandrov LB, Nik-Zainal S, Wedge DC, et al: Signatures of mutational processes in human cancer. Nature 500:415-421, 2013

Supplementary Figure 1



Supplementary Figure 1. Mutation (a) and co-mutation (b) frequencies of cancer genes in n=1,054 MSK-IMPACT sequenced advanced lung adenocarcinoma samples. Panel c displays the distribution of patient characteristics including age, sex and smoking status.

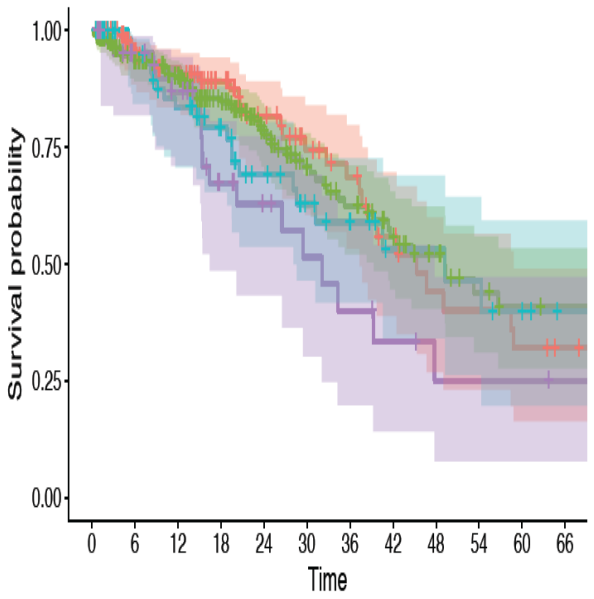
Supplementary Figure 2



Supplementary Figure 2. OncoCast analysis workflow.

Supplementary Figure 3

a

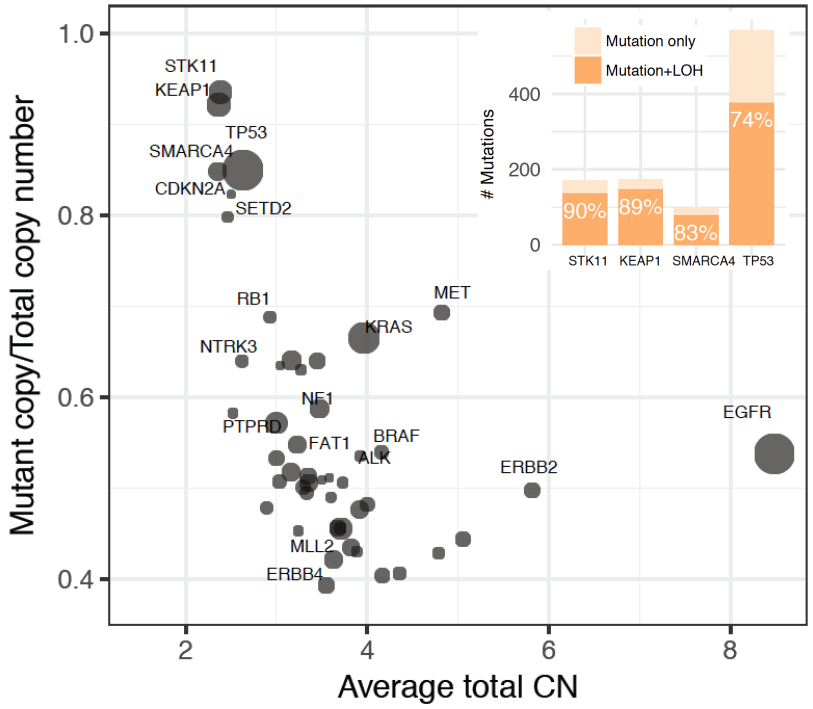


b

	low	intermediate-low	intermediate-high	high
Survival difference				
Hazard ratio	1	1.04	1.22	1.9
95% CI	--	0.67-1.62	0.69-2.14	1.07-3.36
P-value	--	0.85	0.49	0.03
Survival difference adjusted for age, sex, tumor stage				
Hazard Ratio	1	1.31	1.45	1.91
95% CI		0.83-2.06	0.81-2.59	1.06-3.42
P-value		0.24	0.2	0.03
Co-mutation %				
STK11, KEAP1	0	0	0	0.54
KRAS, STK11	0	0	0.26	0.41
KRAS, KEAP1	0	0	0.09	0.41

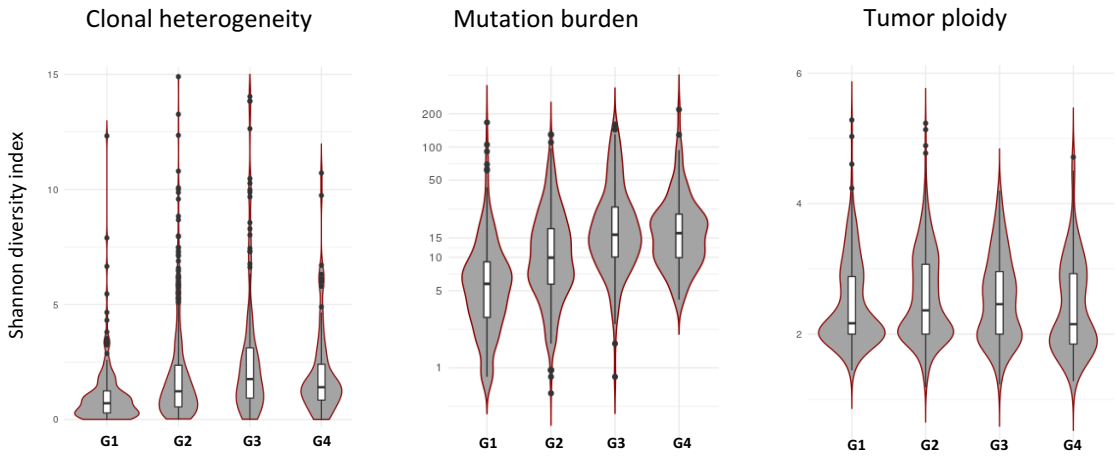
Supplementary Figure 3. Validation of the prognostic model in n=542 TCGA resected lung adenocarcinoma samples.

Supplementary Figure 4



Supplementary Figure 4. Mutant allele copy number analysis of cancer genes.

Supplementary Figure 5



Supplementary Figure 5.Intratumor clonal heterogeneity summarized as Shannon index, overall mutation burden, and tumor ploidy estimated by the FACETS algorithm stratified by the four metastatic lung cancer risk groups.

Supplementary Figure 6

Supplementary Table 1

	Univariate			
	Hazard Ratio	Lower 95% CI	Upper 95% CI	P-value
Genetic risk score*	3.6	2.68	4.83	<2E-16
Targeted	0.58	0.46	0.72	1E-06
Immuno	0.97	0.76	1.23	0.8
Chemo	1.47	1.11	1.96	0.008
	Multivariate			
Genetic risk score *	3.12	2.2	4.41	1E-10
Targeted	0.8	0.62	1.05	0.1
Immuno	0.85	0.67	1.09	0.19
Chemo	1.21	0.89	1.64	0.2

*for 6 unit increase

Supplementary Table 1. Univariate and multivariate Cox regression models evaluating the genetic risk score and treatment types.

Supplementary Table 2

	Univariate			
	Hazard Ratio	Lower 95% CI	Upper 95% CI	P-value
Risk score*	3.60	2.68	4.83	<2E-16
mutation burden (log-scale)	1.28	1.15	1.44	1.48E-05
	Multivariate			
	Hazard Ratio	Lower 95% CI	Upper 95% CI	P-value
Risk score*	3.33	2.38	4.65	2.24E-12
mutation burden (log-scale)	1.47	0.49	3.18	0.33

*for 6 unit increase

Supplementary Table 2. Univariate and multivariate Cox regression models evaluating the risk score and mutation burden.