Appendix 1 Supplementary methods

Sample preparation and DNA extraction

Whole blood samples were collected using a BCT (Streck Inc., Omaha, NE, USA). Plasma was prepared using three centrifugation steps with increasing centrifugal force. After centrifugation, plasma and plasma-depleted whole blood was stored at –80 ℃ until cfDNA extraction. cfDNA was extracted from plasma using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Santa Clarita, CA, USA). Genomic DNA (gDNA) was isolated from blood samples using a QIAamp DNA Mini Kit (Qiagen, Santa Clarita, CA, USA). DNA concentration and purity were quantified using an Infinite M200 Pro NanoQuant (Tecan, Switzerland) and a Picogreen fluorescence assay on a Qubit 4.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Fragment size distribution was measured using a 4200 TapeStation instrument (Agilent Technologies, Santa Clara, CA, USA). An AllPrep DNA/RNA Mini Kit (Qiagen, Santa Clarita, CA, USA) was used to purify gDNA from frozen tissues. After extraction, DNA was quantified and fragmented in the same manner as gDNA from plasma-depleted whole blood, and ≤100 ng of sheared DNA was used for library preparation.

Library preparation

Purified gDNA was sonicated (7 min, 0.5% duty, intensity of 0.1, and 50 cycles/burst) into 150–200 bp fragments using a Covaris S2 (Covaris Inc. Woburn, MA, USA). gDNA and plasma DNA libraries were created using a KAPA Hyper Prep Kit (Kapa Biosystems, Woburn, MA, USA). Briefly, after completing end repair and A-tailing according to the manufacturer's protocol, we performed adaptor ligation at 4 ℃, overnight, using a customized adapter (Integrated Device Technology, San Jose, CA, USA). For the library construction of plasma cfDNA, hybrid selection was performed using three customized baits (LungCancer v1, LiquidSCAN v2-PanCancer, or IVD v1.0, GENINUS, Seoul, Korea, Table S1). Each capture bait targeted 36, 38, and 46 cancer-related genes and covered 340, 117, and 174 kb genomic regions across the human genome.

Detection of somatic mutations

First, all bases were subjected to Phred quality filtering using a threshold Q of 30 and only positions where total depths were above 500× were considered for variant identification. To exclude germline mutations in the analysis, non-reference alleles present at a frequency greater than 1% in the matched white blood cell gDNA were removed. The error suppression method using UMIs was used to distinguish true somatic mutations from PCR and sequencing errors. After applying the error suppression method to the sequencing data, the following selection steps were used to eliminate the remaining sequencing errors: (I) variants not significantly greater than the error found in the matched germline DNA (binomial Bonferroni-adjusted $P<0.01$) were filtered out; (II) variant candidates with a high strand bias (90% if supporting reads 20; Fisher's exact test, $P<0.1$ if supporting reads <20) were removed; (III) if the z-statistic of the variants was not significantly higher than the background error obtained from gDNA (Bonferroni-adjusted $P<0.05$), they were excluded from the analysis.

Finally, the mutation candidates were selected according to the following conditions: Allele frequencies ≥0.15% and alternative allele counts ≥5 were selected. For tissue specimens, somatic variants were identified using different criteria: total depth ≥100× and allele frequency ≥2%. In the case of insertions or deletions, variants with an allele frequency ≥5% were selected. Variants were annotated using VEP (v102) (23) and nonsynonymous variants were used in this analysis.

Clinical variables

Demographic and clinical information were obtained from electronic medical records, including age, sex, body mass index (BMI), and smoking status. Tumors were staged using the eighth edition of the American Joint Committee on Cancer (24) and central location was defined as 'within the inner one-third of the hemithorax by concentric lines arising from the midline' (25).

Regarding COPD, dyspnea was measured using the modified Medical Research Council (mMRC) grade, symptom burden measured using the COPD assessment test (CAT), pulmonary function tests (26,27), and chest CT parameters were collected.

https://dx.doi.org/10.21037/tlcr-23-633

All spirometry tests were performed in a pulmonary function lab, using a Vmax 22 system (SensorMedics, Yorba Linda, CA, USA) according to the American Thoracic Society/European Respiratory Society criteria (26). Absolute values were obtained, and the percentages of predicted values were calculated using a reference equation obtained from a representative South Korean sample (27). All chest CT scans were analyzed using automatic segmentation software (Aview, Coreline Soft, Seoul, Korea) (28,29). We measured whole lung volume at inspiration and the emphysema index (EI), defined as the percentage of lung area with CT attenuation values <–950 HU in the whole lung at inspiration. We also measured the EI of the tumor-located lobe. At the time of blood sampling for cfDNA analysis, white blood cell count and high-sensitivity C-reactive protein (hsCRP) were measured together.

Statistical analysis

To analyze the clinical factors associated with the detection of ctDNA in the study participants, we performed logistic regression analyses for continuous variables (age, BMI, EI, and CRP) and categorical variables (sex, mMRC ≥2, CAT ≥10, FEV1 <50% pred, EI 10%, central location, sequencing panels, and tumor stages). Odds ratios (ORs), 95% confidence intervals (CIs), and p-values were obtained from each analysis. In multivariable logistic regression models (Models 1–5), we used a panel type as an adjusted variable because three different panels were used to generate the mutation data. Variables with P<0.05, in Model 1, were included in the multivariable models (Models 2–4) after forward variable selection. Model 5 was constructed by including variables with P<0.05 in Model 2 adjusted by panel. To estimate the prediction score of ctDNA detection in COPD patients, we used the sum of beta coefficients of significant variables from Model 5 (P<0.05; EI (%), CRP, and tumor stage).

To predict ctDNA detection using the variables, we considered four binary classifying ML models [logistic regression (LR), elastic net logistic regression (EN), random forest (RF), and support vector machine (SV)]. After splitting the dataset into training and test sets within the frame of leave-one-out cross-validation, we selected variables as features for ML models that showed significant association (P<0.1) with the presence of ctDNA mutation in a univariable logistic regression model within each training set. The hyperparameters for EN, RF, and SV models were optimized by using grid search 5 cross-validation for accuracy in each training set. EN model was tuned by alpha from 0.0001 to 100, and L1 ratios between 0.0 and 1. RF model was allowed to have 10 to 1,000 estimators, maximum depth between 6 and 12, minimum samples per leaf between 8 and 18, and minimum samples per split between 8 and 20. SV model was allowed to use either radial or linear kernels, with gamma and C parameters between 0.001 to 100. To evaluate each model, we estimated the area under the receiver operating characteristics (ROC) curve (AUC), accuracy, sensitivity, specificity, and positive predictive value in the test set, and represented the performance of each model using an ROC curve plot. The model with the highest AUC was selected as the best prediction model for the shedder.

The Kaplan–Meier method was used to estimate the overall survival (OS). Data of patients who were alive or those who could not be traced during follow-up were censored for OS at the time they were last known to be alive. Hazard ratios (HRs) and 95% CIs were calculated using the Cox proportional hazards model. All analyses were performed using R 3.6.0, Stata 14.0, and Python 3.8.8.

https://dx.doi.org/10.21037/tlcr-23-633

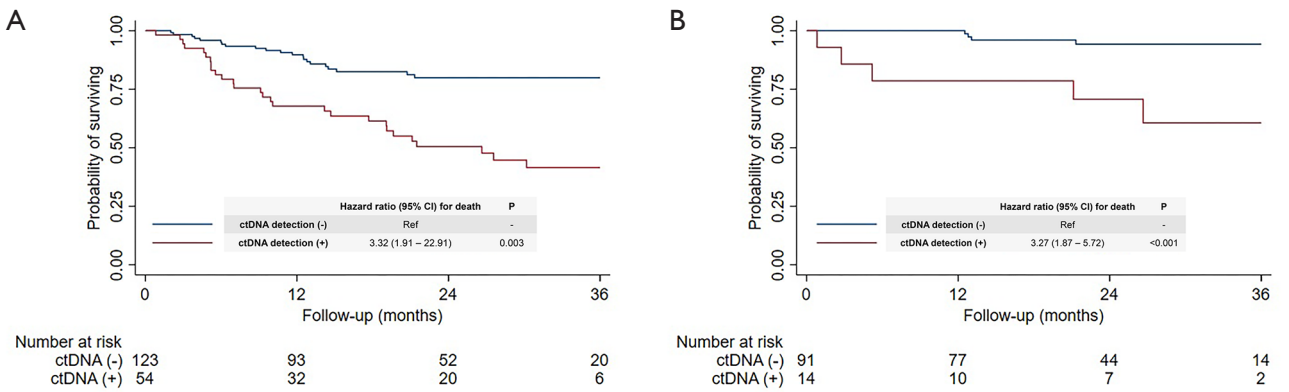

Figure S1 Overall survival according to ctDNA detection in COPD patients with lung cancer of (A) all stages (N=177) and of early stage (N=105).

Table S1 List of cancer-related genes included in targeted deep sequencing panels

Panels	List of genes							
Lung cancer v1	*AKT1*	*ALK*	*ARAF*	*ATM*	*BRAF*	*BRCA1*	*BRCA2*	*CDKN2A*
	EGFR	*ERBB2*	*FGFR1*	*FGFR2*	*FGFR3*	*HRAS*	*IDH1*	*IDH2*
	JAK2	*KEAP1*	*KIT*	*KRAS*	*MAP3K1*	*MDM2*	*MET*	*MYC*
	MYCL	*MYCN*	*NF1*	*NFE2L2*	*NRAS*	*NTRK1*	*NTRK2*	*NTRK3*
	PDGFRA	*PIK3CA*	*PTEN*	*RAF1*	*RB1*	*RET*	*RICTOR*	*ROS1*
	SMARCA4	*STK11*	*TP53*	*TSC1*	*U2AF1*			
LiquidSCAN v2—pan cancer	*AKT1*	*APC*	*BRAF*	*CBFB*	*CDH1*	*CDKN1B*	*CDKN2A*	*CSMD3*
	CTNNB1	*EGFR*	*EPHA5*	*ERBB2*	*ESR1*	*FBXW7*	*FGFR2*	*GATA3*
	GRM8	*HIST1H3B*	*KEAP1*	*KRAS*	*LRP1B*	*MAP2K4*	*MAP3K1*	*MYC*
	NFE2L2	*NRAS*	*NTRK3*	*PIK3CA*	*PIK3R1*	*PPP2R1A*	*PTEN*	*RB1*
	RUNX1	*RYR2*	*SMAD4*	*STK11*	*TBX3*	*TP53*		
IVD v1.0	*AKT1*	*ALK*	*APC*	*AR*	*ATM*	*BRAF*	*BRCA1*	*BRCA2*
	CDH1	*CDKN2A*	*CTNNB1*	*EGFR*	*ERBB2*	*ESR1*	*FBXW7*	*FGFR3*
	GNAS	*HRAS*	*HSPH1*	*KIT*	*KRAS*	*MET*	*MTOR*	*MYC*
	NF1	*NOTCH1*	*NRAS*	*PDGFRA*	*PIK3CA*	*POLE*	*PTEN*	*RB1*
	RET (fusion)	*ROS1(fusion)*	*SMAD4*	*SMARCA4*	*STK11*	*TP53*		

https://dx.doi.org/10.21037/tlcr-23-633

Table S2 Performance of prediction models for ctDNA detection using machine learning according to different variables for the emphysema index

Performance	LR			EN			SV			RF		
	Model a*	Model b	Model c	Model a	Model b	Model c	Model a	Model b	Model c	Model a	Model b	Model c
Accuracy (%)	71.8	71.8	70.3	68.4	65.5	68.0	66.1	71.8	58.3	71.2	70.1	68.6
Specificity (%)	85.4	84.6	83.6	81.3	72.4	75.4	88.6	94.3	78.7	93.5	92.7	91.8
Sensitivity (%)	40.7	42.6	39.6	38.9	50.0	50.9	14.8	20.4	11.3	20.4	18.5	15.1
PPV (%)	55.0	54.8	51.2	47.7	44.3	47.4	36.4	61.1	18.8	57.9	52.6	44.4
AUC	0.767	0.774	0.754	0.650	0.678	0.642	0.557	0.663	0.539	0.719	0.711	0.692

*, for the EI, continuous and binary values were used in model a and model b, respectively, and continuous value of EI of the tumor located in lobes was used in model c. LR, logistic regression; EN, elastic net regression; SV, support vector machine; RF, random forest; PPV, positive predictive value; AUC, area under the receiver operating characteristic curve; EI, emphysema index.

Table S3 Prediction score of the 10th decile group of COPD patients with lung cancer according to Model 5

Sample	ctDNA mutation	EI (%) of total lung	CRP (mg/dL)	Tumor stage	Prediction score	Decile group
COPD_352	Detected	1.098	9.43	3	5.560	10th
COPD_444	Detected	2.556	8.94	3	5.303	10th
COPD_261	Detected	0.067	8.43	3	5.275	10th
COPD_17	Detected	8.031	7.75	4	5.232	10th
COPD_34	Detected	0.054	8.2	3	5.196	10th
COPD_407	Not detected	1.806	6.24	4	5.082	10th
COPD_31	Detected	0.204	4.96	4	4.734	10th
COPD_102	Not detected	7.173	6.95	3	4.335	10th
COPD_393	Detected	1.925	2.42	4	3.749	10th
COPD_227	Detected	0.778	3.9	3	3.661	10th
COPD_190	Not detected	7.626	4.97	3	3.621	10th
COPD_186	Detected	5.823	3.72	3	3.295	10th
COPD_117	Detected	0.030	2.67	3	3.279	10th
COPD_216	Detected	3.620	3.28	3	3.275	10th
COPD_450	Detected	0.708	0.8	4	3.260	10th
COPD_340	Detected	0.601	0.6	4	3.197	10th
COPD_32	Detected	0.283	2.46	3	3.191	10th

Prediction score = −0.060*EI (%) + 0.347*CRP +1.389*Tumor_stage2 + 2.354*Tumor_stage3 + 3.025*Tumor_stage4.

https://dx.doi.org/10.21037/tlcr-23-633

Table S4 Risk of all-cause mortality in COPD patients with lung cancer according to ctDNA detection or VAF (%)

Stage	Unadjusted		Adjusted*	
	HR for death	P	HR for death	P
All stages (N=177)				
ctDNA detection	3.27 (1.87–5.72)	<0.001	1.39 (0.71–2.70)	0.337
VAF (%)	1.04 (1.02–1.05)	<0.001	1.00 (0.98–1.03)	0.687
Stage I, II (N=105)				
ctDNA detection	3.32 (1.91–22.91)	0.003	7.91 (1.55–40.36)	0.013
VAF (%)	1.19 (1.08–1.31)	<0.001	1.25 (1.01–1.56)	0.042
Stage III, IV (N =72)				
ctDNA detection	0.96 (0.51–1.79)	0.886	1.27 (0.63–2.57)	0.511
VAF (%)	1.02 (1.00–1.03)	0.108	1.02 (0.99–1.04)	0.185

*, adjusted for age, smoking (current vs. former), BMI, FEV_1 % pred, emphysema index of total lung (%), CRP, clinical stage of lung cancer, central location, and small cell histology. In a subgroup analysis by early and advanced stages, clinical stage was not adjusted. BMI, body mass index; COPD, chronic obstructive pulmonary disease; CRP, C-reactive protein; ctDNA, circulating tumor DNA; EI, emphysema index; HR, hazard ratio; VAF, variant allele frequency.

https://dx.doi.org/10.21037/tlcr-23-633