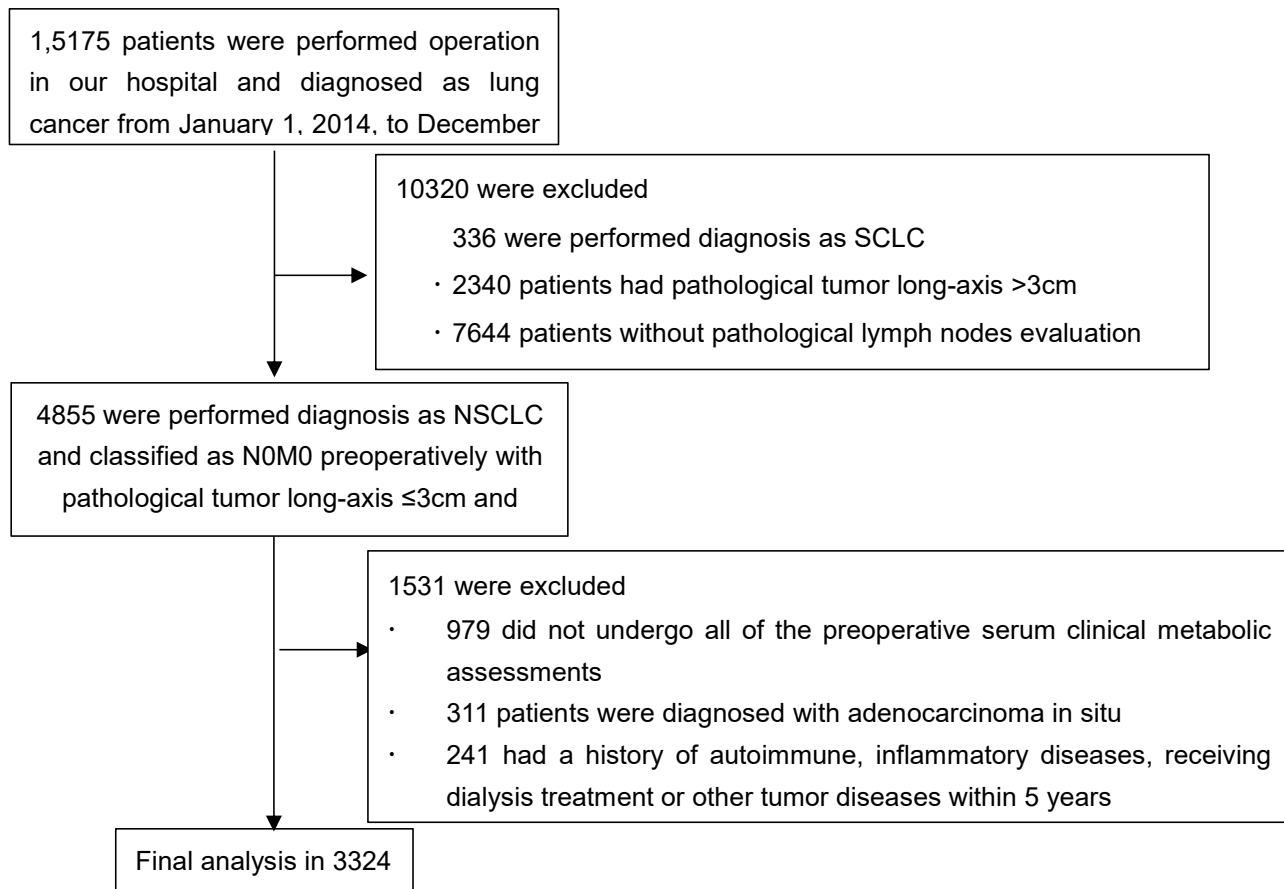
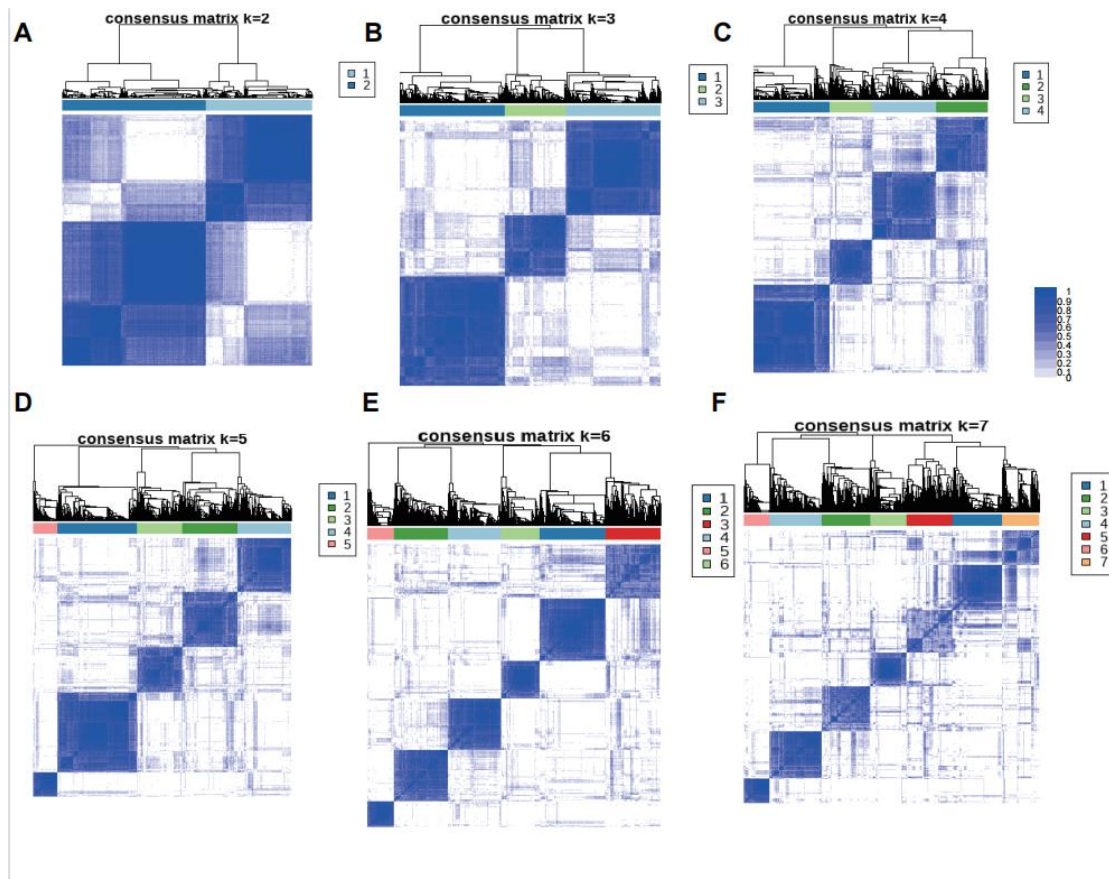


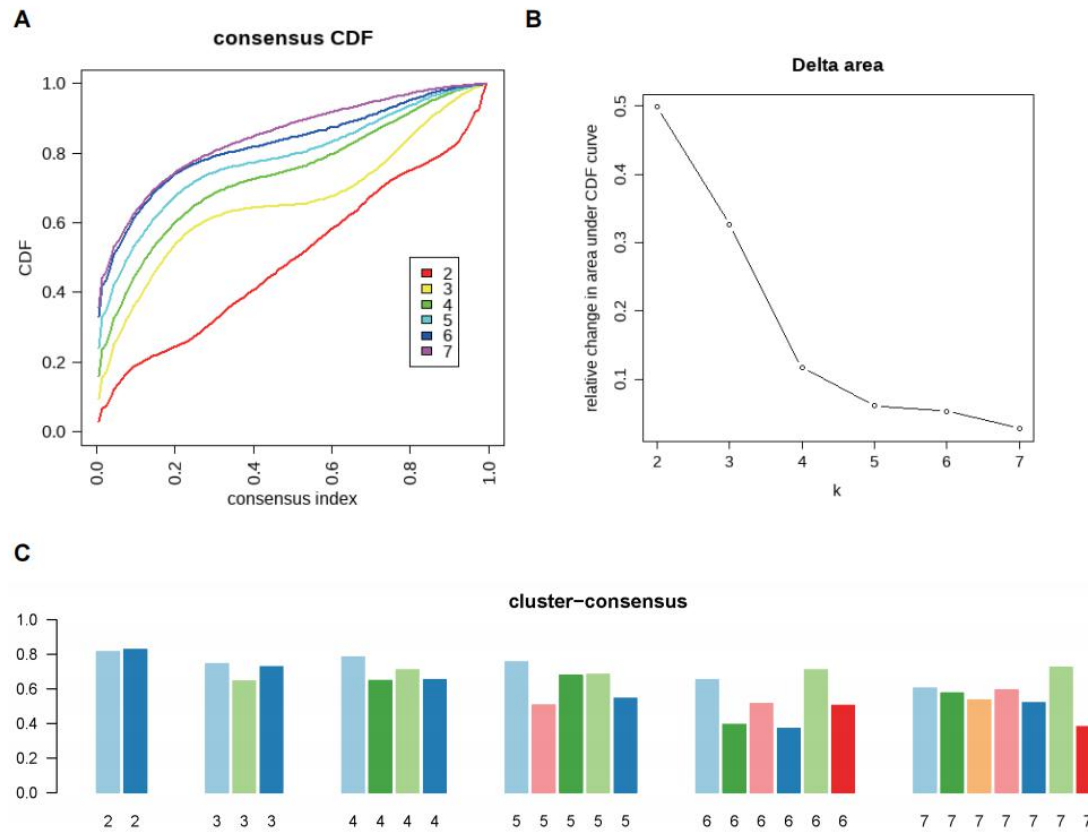
Supplementary Material



Supplementary figure 1. Participant flow diagram of the patients



Supplementary figure 2. Consensus matrix heatmaps using serum metabolic-related measurements in patients with tumor less than 2 cm. (A) K = 2. (B) K = 3. (C) K = 4. (D) K = 5. (E) K = 6. (F) K = 7. The consensus matrix heat maps of K = 2 to K = 7 using 26 serum metabolic-related measurements, including glucose, uric acid, triglycerides, total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, creatinine, adenosine deaminase, lactate dehydrogenase, alanine aminotransferase, aspartate aminotransferase, gamma-glutamyl-transferase, alkaline phosphatase, prealbumin, globulin, albumin, total protein, total bilirubin level, direct bilirubin, indirect bilirubin, carcinoembryonic antigen, cancer antigen, squamous cell carcinoma antigen, progastrin-releasing peptide, the soluble fragment of cytokeratin 19, neuron-specific enolase. (n = 2161). The blue color represents perfect consensus where two individuals always group together, the white color represents perfect consensus where two individuals always group separately, and the blue color scales in between represent ambiguous consensus where two individuals are grouped together in some runs but separately in others.



Supplementary figure 3. Consensus cumulative distribution function and cluster consensus score in patients with tumor less than 2 cm. (A) The lines by colors indicate the cumulative distribution functions (CDF) of the consensus matrix for each K, determined by a histogram of 100 bins. If the CDF reaches an approximate maximum, consensus and cluster confidence are at a maximum at this K (B) The relative change in area under the CDF curve estimated by comparing K and K – 1. For K = 2, there is no K – 1, so the total area under the curve rather than the relative increase is plotted. The relative increases in consensus are used to determine K at which there is an appreciable increase. (C) The bar plot denotes the mean consensus score for different numbers of clusters (K ranges from two to seven) on the basis of 100 repeated re-samplings of 80% of the 2161 small-sized NSCLC patients with tumor less than 2 cm. The bars are grouped by K, which is marked on the horizontal axis. For K = 3, the mean consensus score was 0.78 for cluster 1, 0.76 for cluster 2, 0.79 for cluster 3.

Supplementary Table 1. The characteristics of NSCLC subjects with tumor less than 2 cm by clusters

	cluster 1 (N=891)	cluster 2 (N=497)	cluster 3 (N=773)	q
Female (%)	681 (76.4)	354 (71.2)	256 (33.1)	<0.0001
Smoker (%)	134 (15.0)	114 (22.9)	304 (39.3)	<0.0001
BMI, kg/m2	24.5 (22.5, 26.63)	23.85 (21.9, 26)	24.9 (22.8, 27.08)	<0.0001

Age, year	63 (57, 68)	64 (55, 70)	63 (56, 69)	0.66
Glu mmol/L	5.27 (4.83, 5.89)	4.92 (4.56, 5.43)	5.15 (4.72, 5.76)	<0.0001
UA,μmol/L	287 (245, 336)	263 (221, 309)	345 (293, 403)	<0.0001
TC,mmol/L	5.6 (4.97, 6.33)	4.81 (4.21, 5.4)	4.7 (3.98, 5.26)	<0.0001
TG,mmol/L	1.17 (0.86, 1.67)	0.94 (0.7, 1.38)	1.17 (0.88, 1.64)	<0.0001
LDL_C,mmol/L	3.35 (2.82, 3.89)	2.77 (2.42, 3.24)	2.75 (2.26, 3.24)	<0.0001
HDL_C,mmol/L	1.6 (1.35, 1.82)	1.33 (1.13, 1.58)	1.31 (1.14, 1.52)	<0.0001
Crea,μmol/L	58 (48, 74.05)	75 (63, 84)	75.9 (61, 89)	<0.0001
ADA U/L	10 (8, 12)	10 (8, 12)	8 (7, 11)	<0.0001
LDH, U/L	179 (162, 200)	156 (138, 173)	169 (149, 188)	<0.0001
ALT, U/L	17 (13.25, 23)	16 (11, 20)	20 (15, 28)	<0.0001
AST, U/ml	18 (16, 22)	16 (14, 18)	18 (15, 22)	<0.0001
GGT, UL	17 (13, 23)	14 (10, 19)	20 (15, 31)	<0.0001
ALP, U/L	69 (57, 84)	62 (51.6, 74)	63 (53, 74)	<0.0001
PA, g/L	295 (262, 331.05)	243 (211.7, 273.5)	312.2 (275.5, 351.1)	<0.0001
GLO, g/L	27.3 (24.65, 30.04)	28 (25.23, 30.6)	25.4 (23.05, 28)	<0.0001
ALB,g/L	54.1 (46.11, 57.85)	40.6 (38.16, 43.97)	46.1 (42, 56.3)	<0.0001
TP,g/L	100 (74.62, 100)	69.04 (64.8, 75.1)	71.96 (66.4, 100)	<0.0001
TBIL,μmol/L	12.53 (10.22, 15.69)	10.9 (8.5, 13.9)	16.1 (12.44, 20.82)	<0.0001
IBIL,μmol/L	9.45 (7.52, 11.83)	7.66 (5.9, 10)	11.2 (8.51, 14.9)	<0.0001
DBIL,ng/ml	3.11 (2.38, 3.93)	3.1 (2.48, 4.18)	4.8 (3.7, 6.11)	<0.0001
CEA,ng/ml	1.85 (1.22, 3.12)	1.7 (1.11, 2.73)	1.99 (1.33, 3.15)	<0.0001
SCC,ng/ml	0.98 (0.71, 1.45)	0.85 (0.54, 1.29)	1.15 (0.81, 1.65)	<0.0001
proGRP,pg/ml	30.85 (25.64, 38.65)	29.79 (24.19, 37.95)	31.5 (25.46, 40.23)	0.007
cyfra21-1,ng/ml	2.16 (1.63, 3.02)	1.98 (1.51, 2.78)	2.31 (1.7, 3.1)	<0.0001
CA125,U/ml	9.85 (7.58, 13.79)	10.95 (7.52, 15.32)	9.31 (7.16, 13.16)	0.0001
NSE,ng/ml	13.27 (11.75, 15.11)	11.3 (9.84, 13.2)	12.88 (11.19, 14.86)	<0.0001

Data are median (quartile 1-quartile 3) for continuous variables and number (percentage) for categorical variables. Glu, glucose; UA, uric acid; TG, triglycerides; TC, total cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; Crea, creatinine; ADA, adenosine deaminase; LDH, lactate dehydrogenase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, gamma-glutamyl-transferase; ALP, alkaline phosphatase; PA, prealbumin; GLO, globulin; ALB, albumin; TP, total protein; TBIL, total bilirubin level; DBIL, direct bilirubin; IBIL, indirect bilirubin; CEA, carcinoembryonic antigen; CA125, cancer antigen 125; SCC, squamous cell carcinoma antigen; proGRP, progastrin-releasing peptide; cyfra21-1, the soluble fragment of cytokeratin 19; NSE, neuron-specific enolase. The adjusted P-value (q) was calculated with Benjamini–Hochberg (BH) method to correct the multiple comparisons.

Supplementary Table 2. The association between NSCLC clusters and lymph node metastasis in patients with tumor less than 2 cm.

	Prevalence, %	Unadjusted		Adjusted	
		OR (95%CI)	P value	OR (95%CI)	P value
Cluster 1	5.39 (4.00,7.08)	Reference		Reference	
Cluster 2	11.87 (9.16, 15.04)	2.13 (1.52, 3.01)	<0.0001	2.19 (1.54, 3.17)	<0.0001
Cluster 3	9.06 (7.13, 11.30)	1.46 (1, 2.13)	0.052	1.98 (1.32, 2.97)	0.0009

Data are prevalence rate (95% CIs) and odds ratio (95% CI). Odds ratio (OR) and 95% confidence interval (CI) was evaluated using multivariable logistic regression models.

Adjusted model was adjusted for age, gender, body mass index, smoking status.