Supplementary Information for

Hexokinase 2 discerns a novel circulating tumor cell population associated with poor prognosis in lung cancer patients

Liu Yang[a,1], Xiaowei Yan[b,1], Jie Chen[c,1],Qiong Zhan[d,1], Yingqi Hua[a,1], Shili Xu[e], Ziming Li[f], Zhuo Wang[g,h], Yu Dong[c], Dongqing Zuo[a], Min Xue[g], Yin Tang[b], Harvey R. Herschman[e], Shun Lu[f,2], Qihui Shi[g,h,2], and Wei Wei[b,e,2]

[a]Shanghai Bone Tumor Institute and Department of Orthopedics, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [b]Institute for Systems Biology, Seattle, WA, USA; [c]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China; [d]Department of Oncology, Huashan Hospital, Fudan University, Shanghai, China; [e]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, CA, USA; [f]Shanghai Lung Cancer Center, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China; [g]Shanghai Key Laboratory of Medical Epigenetics (Ministry of Science and Technology), Institutes of Biomedical Sciences, Fudan University, Shanghai, China; [h]Key Laboratory of Whole-period Monitoring and Precise Intervention of Digestive Cancer (SMHC), Minhang Hospital & AHS, Fudan University, Shanghai, China.

[1]These authors contributed equally to this work.

[2]Correspondence: S.L. (shunlu@sjtu.edu.cn); Q.H.S. (qihuishi@fudan.edu.cn); and W.W. (wwei@systemsbiology.org)

I. Supplementary Materials and Methods

II. Supplementary Figures S1-S17

III. Supplementary Tables S1-S7

IV. Supplementary Datasets S1-S2

I. Supplementary Materials and Methods

Cell lines and reagents. All the cancer cell lines used in this study were obtained from ATCC and routinely maintained in RPMI-1640 medium (Life Technologies) containing 1× Penicillin-Streptomycin-Glutamine (Gibco) and 10% FBS (Gibco) in a humidified atmosphere of 5% CO_2 and 95% air at 37ºC. (APC)-conjugated CD45 (clone HI30), Alexa Fluor 488-conjugated goat-anti-rabbit secondary antibody (#A11008) and MitoTracker™ Green FM (#M7514) were purchased from ThermoFisher Scientific. PE-conjugated anti-cytokeratin 7/8 antibody was purchased from BD Biosciences (#347204). Anti-HK2 primary antibody was purchased from Abcam (#ab209847). Anti-Vimentin antibody was purchased from CST (#5741S). DAPI was purchased from Beyotime Biotechnology. Hank's balanced salt solution (HBSS, no calcium, no magnesium, no phenol red) was purchased from Gibco. All primers were synthesized by Genewiz and listed in *SI Appendix*, Table S7.

Fabrication of microwell chip. The microwell chip was fabricated in poly(dimethylsiloxane) (PDMS) using standard microfabrication soft-lithographic techniques. A replicate for molding the PDMS was obtained by patterning a silicon wafer using photoresist SU-8 2050. The PDMS pre-polymer (Sylgard 184, Dow Corning) was mixed in a ratio of 10:1, and subsequently casted on this lithographically patterned replicate. After curing at 80 °C for 2 h, the PDMS component was separated from the replicate. The diameter and depth of the microwells are 30 μm and 20 μm, respectively. A chip containing 28,000 microwells was used to measure peripheral blood and cerebrospinal fluid samples, and a chip containing 112,000 microwells was used to measure pleural effusion samples.

Patient information and sample collection. The clinical study was approved by the institutional ethics review committees at the Shanghai Chest Hospital and Huashan Hospital and was performed according to the Declaration of Helsinki Principles. Peripheral blood and pleural effusion samples were obtained from LUAD patients in Shanghai Chest Hospital with written informed consent. Cerebrospinal fluid samples were obtained from LUAD patients in Huashan Hospital with written informed consent. All patients were treatment-naïve and at stage III or IV. Fifty patients who contributed peripheral blood samples were enrolled from June 2018 to December 2018 and from April 2020 to November 2020 and followed until November 2020. The tumors were staged according to the 8th edition of the international tumor-node-metastasis (TNM) system. The healthy volunteers had no known illness or fever at the time of blood draw, no history of malignant disease, and were >30 years old.

Identification of CTCs in peripheral blood samples. Fresh blood samples (5 mL) were drawn and preserved in TransFix/EDTA Vacuum Blood Collection Tubes and delivered to the lab within 4 hours. Blood samples were initially centrifuged at 500 g for 5min. The supernatant was discarded and the cell pellet was re-suspended in an equivalent volume of HBSS and mixed with 25µL CTC enrichment antibody cocktail (RosetteSep^{TM} CTC Enrichment Cocktail Containing Anti-CD36, STEMCELL Technologies) at room temperature for 20 min. 15 mL HBSS with 2% FBS (Gibco) was then added and the samples were mixed well. The mixture was carefully added along the wall of the Sepmate tube (SepMate^{TM}-50, STEMCELL Technologies) after adding 15 mL density gradient liquid (Lymphoprep^{TM}, STEMCELL Technologies) into the tube through the middle hole. After centrifuging at 1200 g for 20 min, the topmost supernatant (~10 mL) was discarded, and the remaining liquid (~10 mL) above the barrier of the Sepmate tube was rapidly poured out into a new centrifuge tube. After centrifuging at 600 g for 8 min, the supernatant was removed and 1 mL of red blood cell lysing buffer (BD Biosciences) was then added for 5 min to lyse red blood cells. After centrifuging at 250 g for 5 min, the nucleated cell pellet was re-suspended in HBSS. The cell suspension was then applied onto the 3% BSA (Sigma)-treated microwell chip. A 10 min waiting period was allowed for the cells to sit down in the 28,000 microwells. The chip was sealed with a porous Isopore^{TM} polycarbonate membrane (Merck Millipore, #TSTP04700) to avoid cell loss during subsequent on-chip staining. After cell fixation (2% PFA, 10 min) and permeabilization (0.5% Triton X-100, 15 min), blocking solution consisting of 3% BSA and 10% Normal Goat Serum was applied to the chip for 1 h, followed by incubation with APC-conjugated anti-CD45 antibody (mouse), PE-conjugated anti-CK 7/8 antibody (mouse) and anti-HK2 antibody (rabbit) in PBS overnight at 4 ºC. After extensive washing with PBS, cells on the chip were treated with Alexa Fluor 488-conjugated goat-anti-rabbit secondary antibody in PBS for 1 h and DAPI for 10 min followed by washing with PBS. ImageXpress Micro XLS Wide field High Content Screening System (Molecular Devices) scanned the chip and imaged all cells in bright field and four fluorescent colors (CD45: CY5; HK2: FITC; CK: TRITC; Nucleus: DAPI). A computational algorithm analyzed the images and identified HK2^{high} and CK^{pos} cells based on the cut-offs generated from HK2 and CK fluorescence intensity of CD45^{pos} leukocytes in the samples.

Identification of CTCs in malignant pleural effusion (MPE) and cerebrospinal fluid (CSF) samples. For CTC identification in MPE, 5 mL of an MPE sample was first filtered by a membrane with a pore size around 100 μm, followed by centrifuging at 500 g for 5 min to separate cell pellets. 1 mL of red blood cell lysis buffer (BD) was then added to lyse red blood cells for 5 min. After centrifuging at 500 g for 5 min, the nucleated cell pellet was re-suspended in and washed with HBSS. After cell counting, up to 500,000 cells were applied onto a 112,000-well microwell chip for cell fixation, permeabilization and immunostaining as described above. For CTC identification in CSF, CSF samples were centrifuged at 500 g for 5 min to separate cell pellets and resuspended in HBSS, followed by processing with the microwell chip. No enrichment step was included, due to the limited number of cells present in the CSF samples.

CTC recovery and duplicate CTC measurements. Spike-in experiments were performed for assessment of CTC recovery, assay sensitivity, and reproducibility. Different numbers (5, 15, 30, and 50) of HCC827 and H1975 cells were pre-stained with DiI (Beyotime, #C1036) and spiked into 5 mL of blood samples from healthy donors by a micromanipulator. The blood samples were then processed by a single operator according to the protocol described above, including CTC enrichment and on-chip staining step (APC-conjugated anti-CD45 antibody). All cells on the chip were imaged and the spiked cancer cells were identified and counted for computing the cell recovery. Each spike-in experiment was repeated five times. For evaluating the reproducibility of CTC measurement, a total of 7 samples were drawn in duplicate from LUAD patients. All duplicate tube measurements were performed by the same operator.

Single-cell manipulation and *EGFR* mutation detection. Individual candidate CTCs in the microwells were retrieved using a XenoWorks Micromanipulator and trimethylchlorosilane (TMCS)-treated micropipettes, and then transferred into PCR tubes (Axygen). The genome amplification of the retrieved single cells was conducted with the MALBAC® Single Cell Whole Genome Amplification (WGA) Kit (Yikon Genomics). For detection of *EGFR*L858R mutation, PCR reaction was performed using the primers listed in the *SI Appendix*, Table S7. The PCR reaction buffer consisted of 12.5μL 2×Ex Taq DNA polymerase mix (Vazyme Biotech), 10μM forward primer, 10μM reverse primer, and 0.2μL whole genome amplified DNA. The PCR reaction was conducted as follows: 95 ºC for 3 min, followed by 30 cycles (95 ºC for 30 s, 60 ºC for 30 s, and 72 ºC for 30 s), followed by a final extension at 72 ºC for 5min. The PCR products were analyzed by Sanger sequencing (Genewiz, Suzhou, China).

Single-cell whole-genome sequencing. We used 22 pairs of primers targeting 22 loci on different chromosomes to evaluate the WGA coverage of single cells (see *SI Appendix*, Table S7 for primer sequence design) (1). WGA products with 18out of 22 loci amplified were used for subsequent whole-genome sequencing (WGS) library construction. For library construction, WGA products were firstly digested with dsDNA fragmentase (New England Biolabs) and 300-500bp fragments were retrieved with Agencourt AMPure XP beads (Beckman Coulter). Around 100 ng of DNA fragments were used as input for sequencing library construction. DNA fragment repair and library adaptor ligation were performed using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (New England Biolabs), in accordance with the manufacturer's protocol. WGS libraries were amplified using NEBNext® Ultra™ II Q5® Master Mix and index primers (New England Biolabs). Library purification was performed with TIANgel Midi Purification Kits (TIANGEN Biotech). The concentrations of purified fragmented DNA or libraries were quantified with Qubit dsDNA HS Assay Kits (Invitrogen). Libraries were analyzed by Illumina HiSeq X Ten platform with 150 PE at 0.1X coverage (Genewiz, Suzhou, China).

Copy number determination and segmentation from whole-genome sequencing data. Sequencing reads were aligned to the major chromosomes of human (hg19) using BWA (version 0.7.10-r806) with default options (2). SAMtools (version 1.31) was used to mark and remove PCR duplicates (3). To reduce WGS biases, the sequence depths of tumor cells were normalized by GC contents and mappability with 500-kb windows. The diploid regions were determined using HMMcopy (4). Similar copy numbers in adjacent chromosome regions were merged in DNAcopy package to get CNV regions (5).

Processing 10X Genomics single-cell RNA-seq data. Cells were collected from an MPE sample of a LUAD patient. Red blood cell lysis and leukocyte depletion (EasySep Human CD45 Depletion Kit II #17898, STEMCELL Technologies) of MPE were performed before processing with 10X Genomics (performed by CapitalBio Technology, China). RNA-seq profiles of 10,026 single cells were obtained using the Cell Ranger pipelines (version 3.1.0) and the provided reference (refdata-cellranger-GRCh38-3.0.0). Appropriate data quality was confirmed before data analysis (*SI Appendix*, Fig. S17). Clusters of cells were identified using the k-means clustering algorithm as implemented in the Cell Ranger pipeline. Mitochondrial count percentage was calculated for each cell (*SI Appendix*, Fig. S12). A cluster was defined as stressed or dying if the average mitochondrial count percentage was more than 30%. One special cluster was close to dying clusters in the UMAP and the expressed genes number of that cluster was small, indicating that this cluster contained cells of poor quality (*SI Appendix*, Fig. S12). After removing these stressful/dying and poor-quality cells, five main cell types, including malignant, mesothelial, myeloid, lymphoid, and erythroid cells, were identified based on the clustering and the UMAP dimension reduction (Fig. 4A).

CNV inference using the single-cell transcriptome data. For each of those 10,026 single cells from 10X Genomics RNA sequencing, we inferred an estimated CNV profile using R package CONICS over 38 chromosome arms (*SI Appendix*, Fig. S13), where the CNV profile was simulated by the posterior probabilities based on a two-component Gaussian mixture model (6). We first unbiasedly used all 44 autosome arms to infer CNV profiles; however, only 38 of these analyses were able to fit 2-component mixture models. The clustering results showed that all 10,026 single cells can be divided clearly into two clusters, one of which represented malignant cells whose inferred CNV profiles are drastically different from those of benign cell types in the other cluster (*SI Appendix*, Fig. S13).

Calculation of EMT scores using single-cell transcriptome data. The Molecular Signatures Database (MSigDB v7.0, Broad Institute) was used to calculate EMT scores through gene set enrichment analysis (7). R package fgsea was applied to the gene set HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION to calculate a normalized enrichment score for each CTC as its EMT score, with 1000 permutations and the rank list of genes generated as follows: Only those genes observed in at least one CTC were considered. A reference cell was created with the expression of each gene as the average expression of this gene across all the tumor cells. To rank genes for each cell with respect to the reference cell, log2_Ratio_of_Classes was used. Only cells with FDR<0.05 were considered significantly enriched in either epithelial or mesenchymal phenotypes.

Analysis of differentially expressed genes. For the 10,026 single cells from 10X Genomics RNA sequencing, we used the normalized UMI counts for their expression levels. The normalization scaled the raw UMI counts for each barcode by multiplying the ratio of the median UMI sum of all barcodes to the sum of UMI counts of this barcode, assuring that all normalized UMI sums of barcodes have the same as that of the median barcode. After normalizing and log10-transformation of UMI counts, to be consistent with our pan-CK (anti-CK7/8) staining used for CTC identification, the CK (cytokeratin) level in each of the tumor cells was defined as the sum of *KRT7* and *KRT8* expression levels in the cell. CK^{low} cells were defined as cells with CK expression levels lower than the 10% probabilistic quantile, and CK^{high} cells were defined as cells with CK levels higher than the 90% probabilistic quantile. Any CTC was considered to be mesenchymal if its EMT score (see above Calculation of EMT scores using single-cell transcriptome data**)** was positive and FDR < 0.05, and epithelial if its EMT score was negative and FDR <0.05. In either epithelial or mesenchymal population, R package LIMMA was used to compare expressions of CK^{high} cells versus CK^{low} cells and to calculate the fold change and p value for each gene. Only those genes with p values < 0.05 were considered as differentially expressed genes between CK^{high} cells and CK^{low} cells. Top 100 genes up-regulated in CK^{high} cells and down-regulated in CK^{high} cells (i.e. up-regulated in CK^{low} cells) in terms of fold changes were respectively selected and compared in Figure 5.

Gene set variation analysis. To compare the molecular signatures differentially enriched in CK^{high} cells or CK^{low} cells in either epithelial or mesenchymal population, R package GSVA was applied to the CK^{high} and CK^{low} cells to derive a GSVA enrichment score for each cell under a certain gene set (8). Gene sets obtained from the Molecular Signatures Database (MSigDB v7.0, Broad Institute) include the H (hallmark gene sets), C2 (curated gene sets), and C5 (GO gene sets). Normalized and log10-transformed UMI counts were input as the expression matrix. Since GSVA would discard genes with constant expression values across all cells, only those genes with their coefficients of variation (CVs) larger than 1 were selected. We obtained a very similar matrix of GSVA enrichment scores using all the genes as well. All other parameters were set to their default values, except for the sizes of gene sets. We set the minimum size (min.sz) as 10 and the maximum size (max.sz) as 900. After deriving the gene-set-by-cell matrix of GSVA enrichment scores, we used R package LIMMA to compare GSVA enrichment scores of CK^{high} cells with those of CK^{low} cells and to determine if a gene set had a significantly differential enrichment between CK^{high} and CK^{low} cells for a given gene set, via score difference and p-value.

Analysis of LUAD RNA-seq profiles from TCGA. HTSeq-counts profiles of 533 primary LUAD tumor samples and 2 recurrent tumor samples were downloaded from the Genomic Data

Commons (https://portal.gdc.cancer.gov/). Raw read counts were first normalized to counts per million (CPM) using R package edgeR. A reference profile was then created by averaging over all 535 samples. For each of these 535 samples, the EMT score was calculated via gene set enrichment analysis in the same way described above using R package fgsea and the hallmark EMT gene set from MSigDB. The reference profile was created with the expression of each gene as the average expression of this gene across all LUAD samples. The sample was considered to be mesenchymal if its enrichment score was positive and FDR <0.05, and epithelial if its enrichment score was negative and FDR <0.05.

Statistical analysis. Statistical analyses were performed using GraphPad PRISM 8 (GraphPad Software, Inc) unless noted elsewhere. Statistical significance between two groups was compared using the two-tailed Mann-Whitney test with $p < 0.05$ as the significance threshold. Alpha level was corrected by Bonferroni correction when multiple groups were compared. The Kaplan–Meier method was used to estimate the survival rate, along with a log-rank statistical test (two-sided) comparing the survival distribution.

Supplementary References

1. Li Z*, et al.* (2019) Liquid biopsy-based single-cell metabolic phenotyping of lung cancer patients for informative diagnostics. *Nat Commun* 10(1):3856.
2. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
3. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
4. Benjamini Y & Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40(10):e72.
5. Olshen AB, Venkatraman ES, Lucito R, & Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557-572.
6. Muller S, Cho A, Liu SJ, Lim DA, & Diaz A (2018) CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* 34(18):3217-3219.
7. Subramanian A*, 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(43):15545-15550.
8. Hanzelmann S, Castelo R, & Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7.

II. Supplementary Figures

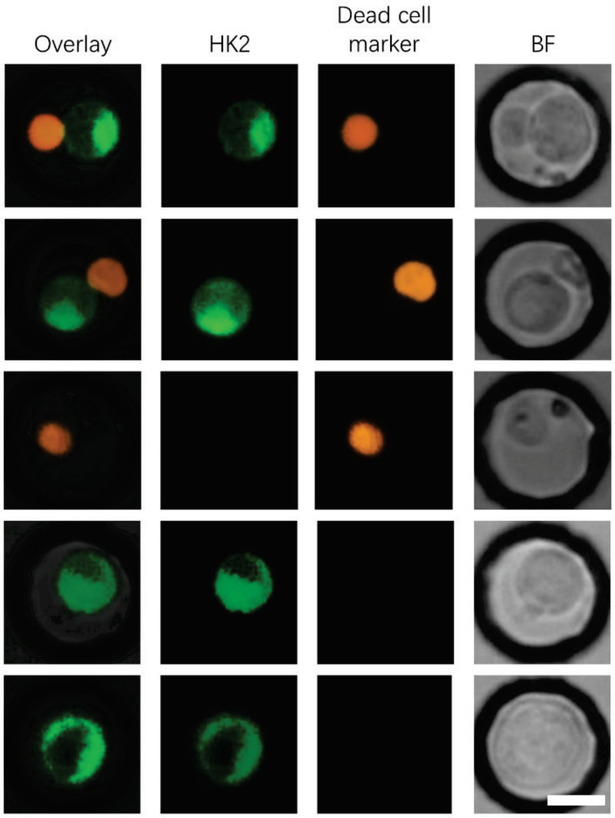


Figure S1. Dead H1975 lung cancer cells are absent of HK2 fluorescence signals. The dead cell marker is from the Live/Dead Viability/Cytotoxicity kit (ThermoFisher, Catalog L3224). Scale bar: 15 μm.

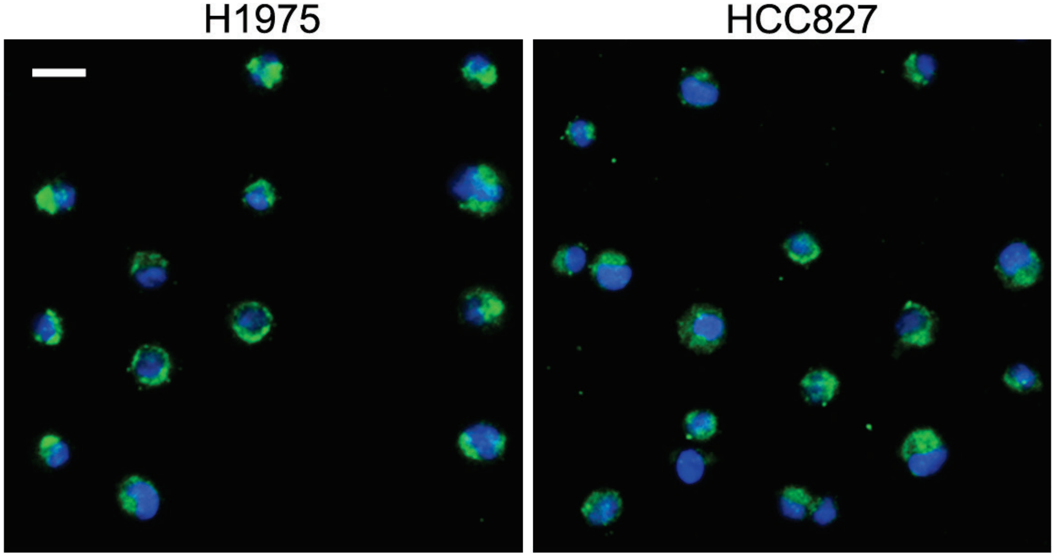


Figure S2. Representative fluorescence images of H1975 and HCC827 cells stained with anti-HK2 and DAPI. Scale bar: 20 μm.

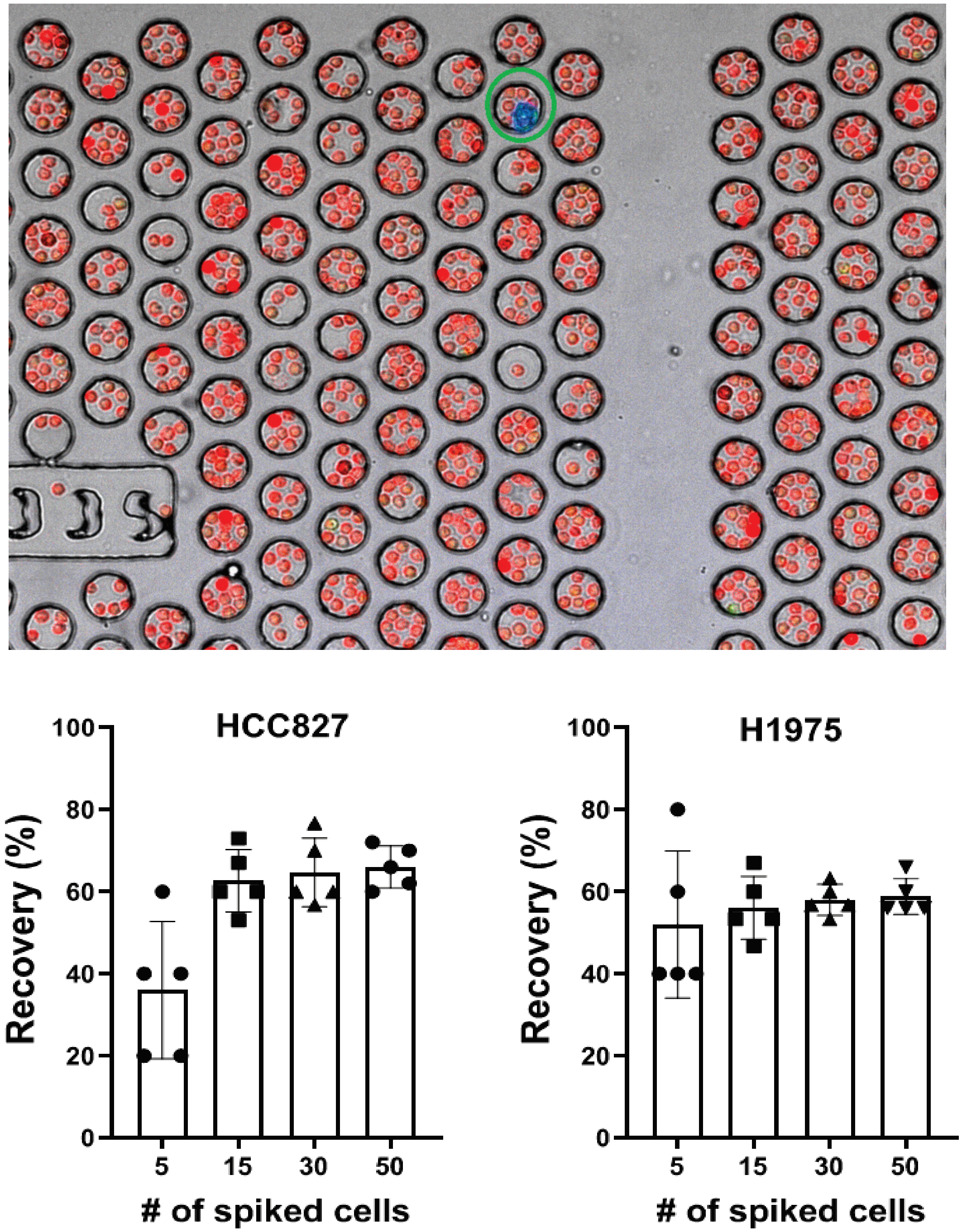


Figure S3. Spike-in experiments for assessment of CTC recovery and assay reproducibility. Defined numbers of HCC827 or H1975 cells were pre-stained with DiI (Beyotime, #C1036) and spiked into 5 mL of blood samples from healthy donors by a micromanipulator, followed by CTC enrichment and on-chip staining with APC-conjugated anti-CD45 antibody as described in the experimental section. Top, the overlaid image shows a spiked H1975 cell (blue) and leukocytes (red) on the microwell chip. Bottom, calculated cancer cell recovery at different numbers of spiked cells. Mean ± SD was shown in the plots.

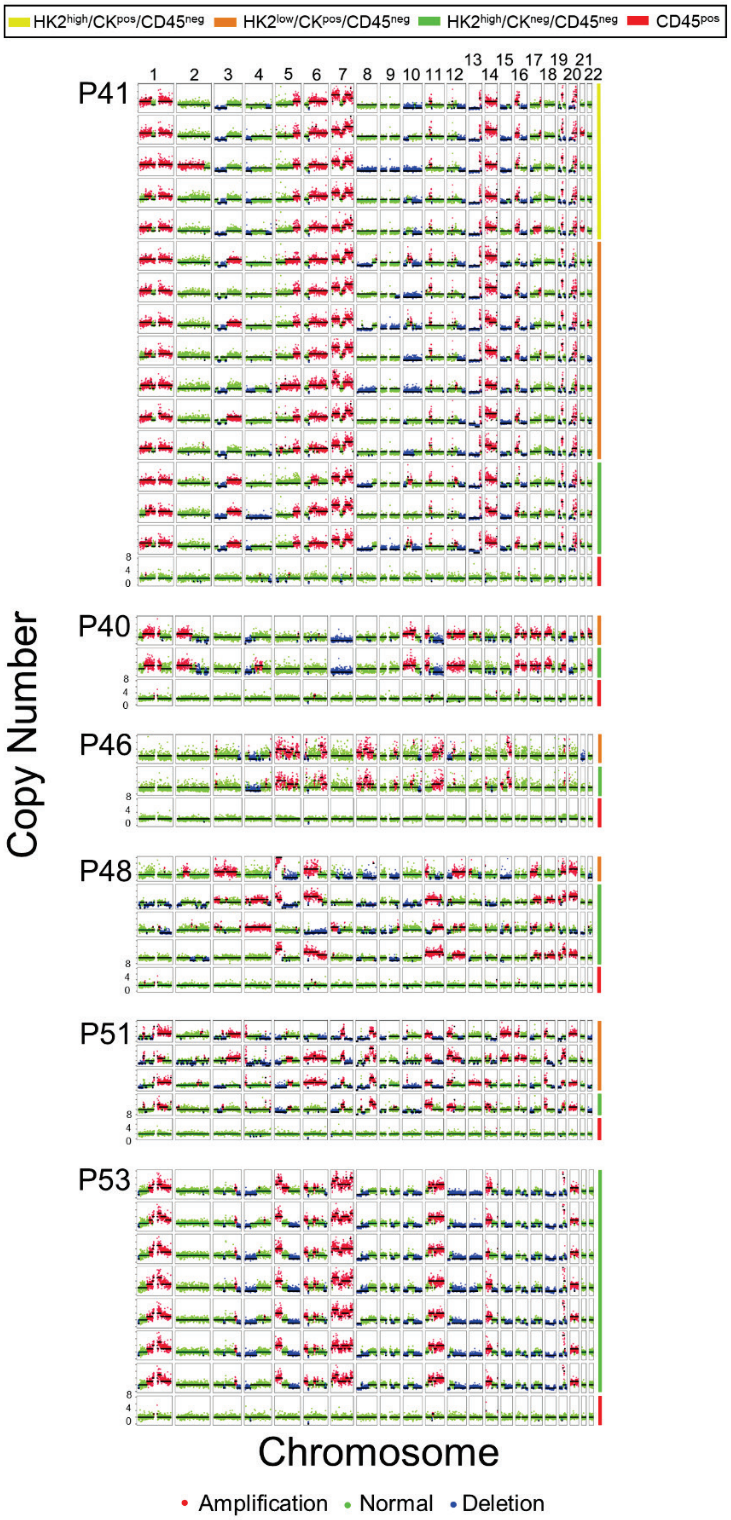


Figure S4. Single-cell CNV profiles across the autosomes of randomly selected CTCs and leukocytes from blood samples of six patients, highlighting their genome-wide similarity independent of CK expression. The CTC subtypes and leukocytes are color-coded to the right with annotations above.

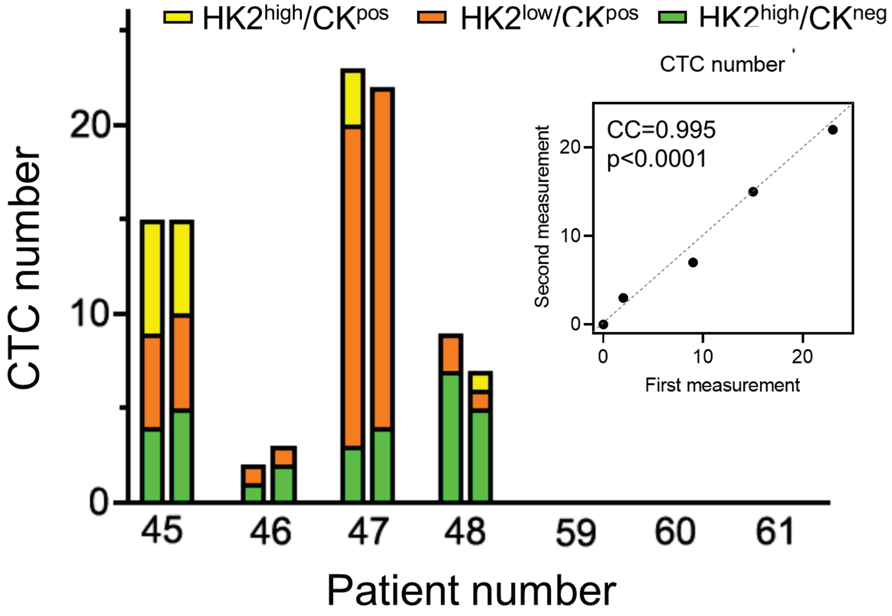


Figure S5. Reproducibility of CTC analysis of 7 blood samples drawn in duplicate from LUAD patients. CTC counts and subtyping of seven blood samples for two independent duplicate measurements were shown in the bar graph. No CTCs were detected in P59, P60, and P61. Pearson correlation coefficient (CC = 0.995) was calculated between two duplicate measurements across seven patients.

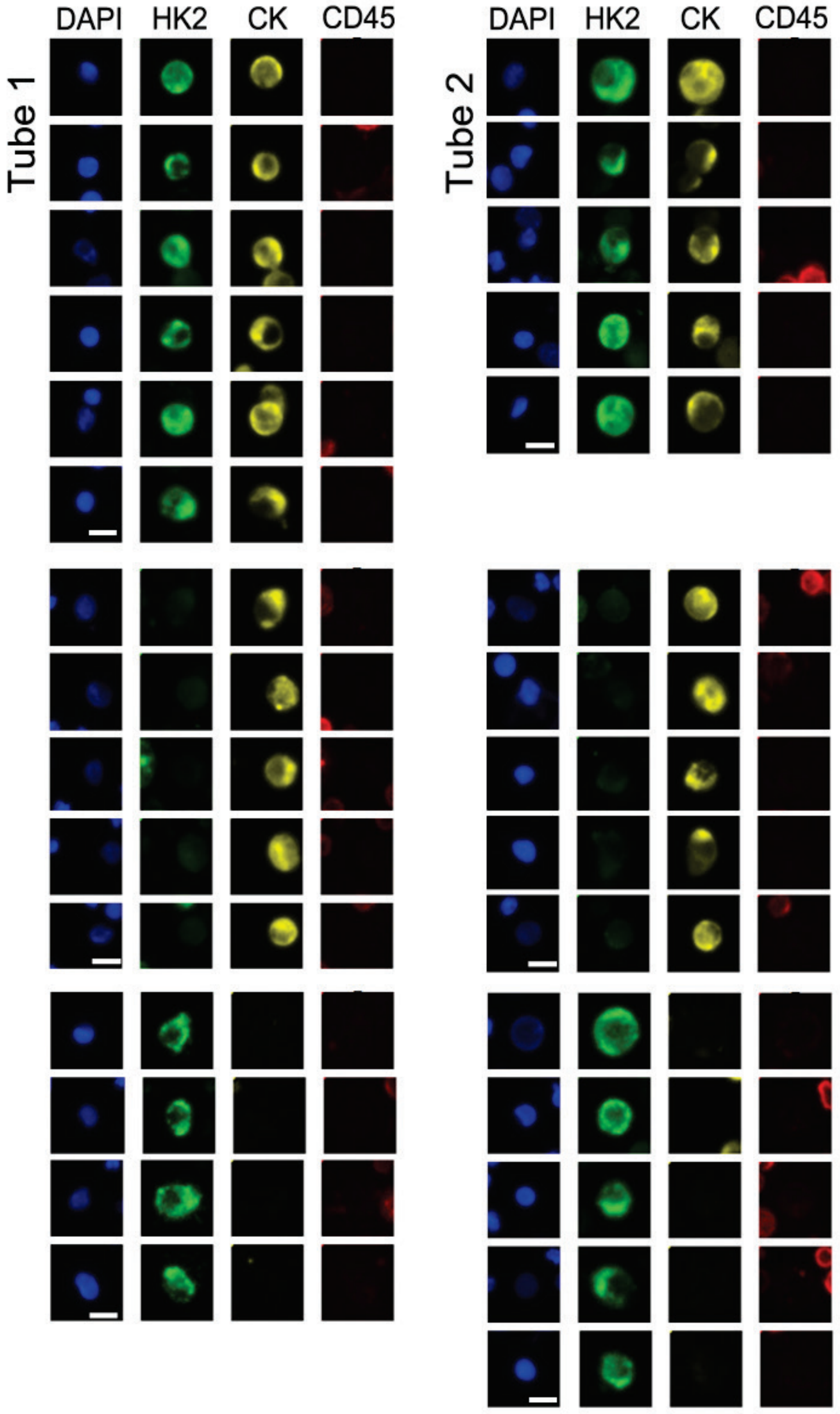


Figure S6. Duplicate CTC detection in the peripheral blood sample from P45. Each tube contained 5 mL of the blood sample and two duplicate tubes were measured. A total of 15 CTCs were detected in both tube 1 and tube 2, and the percentage of CK^{neg} CTCs were 26.7% and 33.3% in tube 1 and tube 2, respectively. Scale bar: 10 μm.

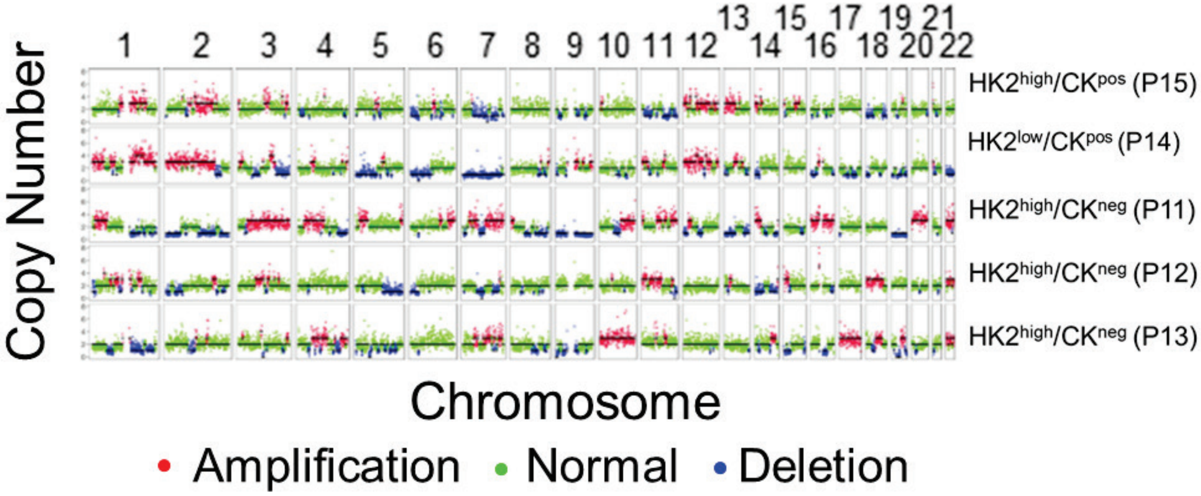


Figure S7. Single-cell CNV profiles across the chromosomes of randomly selected CTCs from blood samples of five patients.

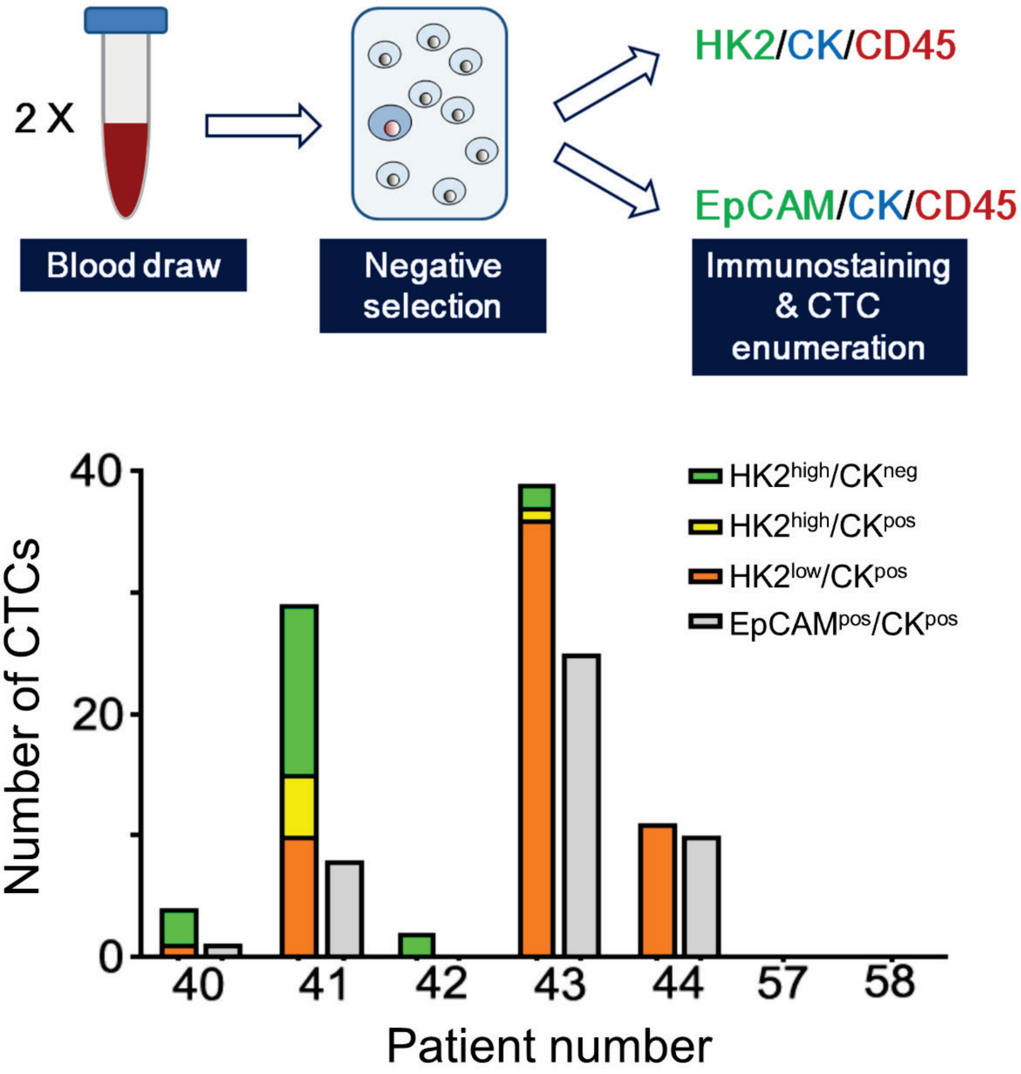


Figure S8. Comparison of CTC detection between the HK2 assay described in this paper and the EpCAM/CK-based, CellSearch-like strategy. Top, schematic description of the strategy. Two duplicate blood samples were drawn from a patient followed by the negative selection for CTC enrichment according to the experimental section in this paper using RosetteSep^{TM} CTC Enrichment Cocktail Containing Anti-CD36 (STEMCELL Technologies). The enriched cells from the two samples were stained for HK2/CK/CD45/DAPI and EpCAM/CK/CD45/DAPI and enumerated, respectively. Bottom, Comparison of CTC counts and subtyping between the two detection strategies across 7 LUAD patients. In duplicate 1, CTCs were identified by the criterion defined in this paper. In duplicate 2, CTCs were identified by the criterion used in the CellSearch system, namely as EpCAM^{pos}/CK^{pos}/CD45^{neg}/DAPI^{pos}.

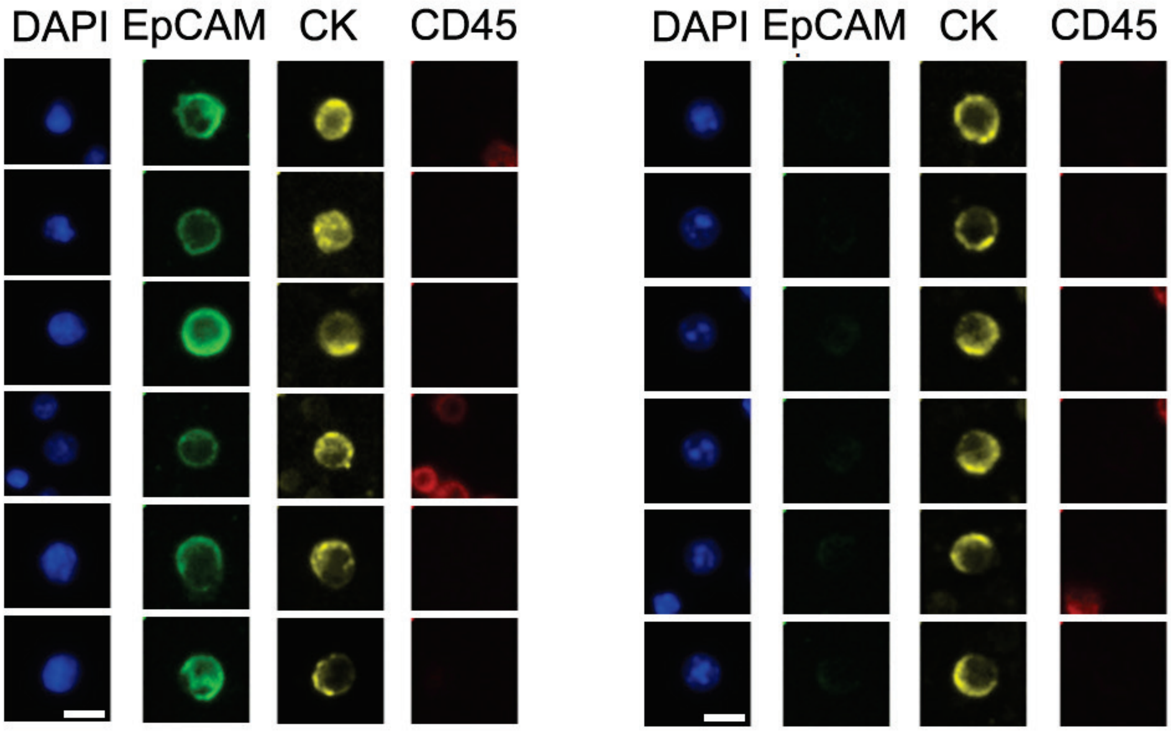


Figure S9. Representative fluorescence images of EpCAM^{pos}/CK^{pos}/CD45^{neg} CTCs and EpCAM^{neg}/CK^{pos}/CD45^{neg} CTCs in the blood sample from P41. Scale bar: 10 μm.

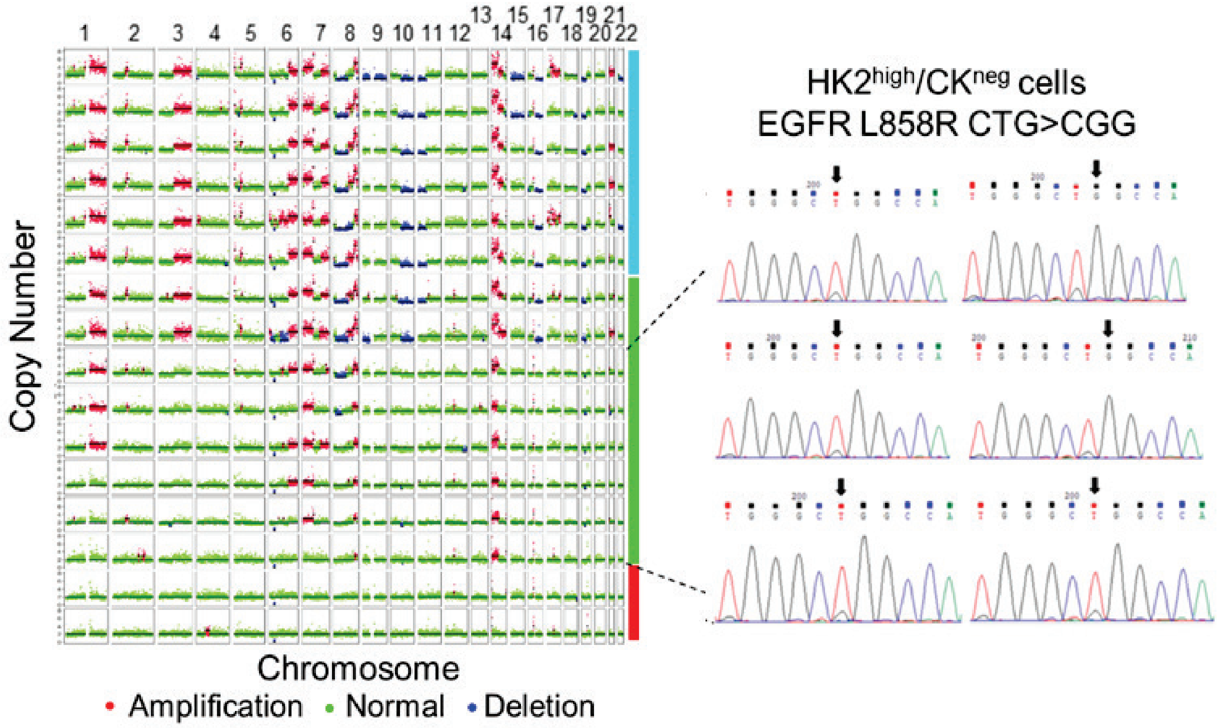


Figure S10. Single-cell CNV profiles of randomly selected CTCs and leukocytes from the MPE sample of P25. HK2^{high}/CK^{pos}/CD45^{neg} and HK2^{high}/CK^{neg}/CD45^{neg} CTCs, as well as leukocytes, are color-coded by cyan, green and red bars to the right, respectively. All these CTCs were detected to harbor the same *EGFR* L858R oncogenic driver mutation as their primary tumor lesion. The representative Sanger sequencing results of selected HK2^{high}/CK^{neg}/CD45^{neg} CTCs are shown to the right.

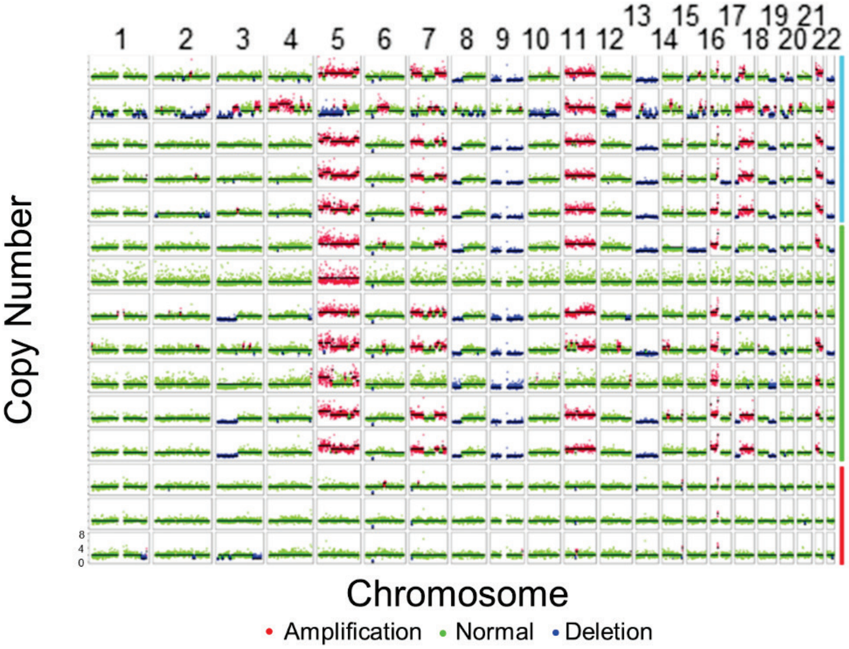


Figure S11. Single-cell CNV profiles of randomly selected CTCs and leukocytes from the MPE sample of P27. HK2^{high}/CK^{pos}/CD45^{neg} and HK2^{high}/CK^{neg}/CD45^{neg} CTCs, as well as leukocytes, are color-coded by cyan, green and red bars to the right, respectively.

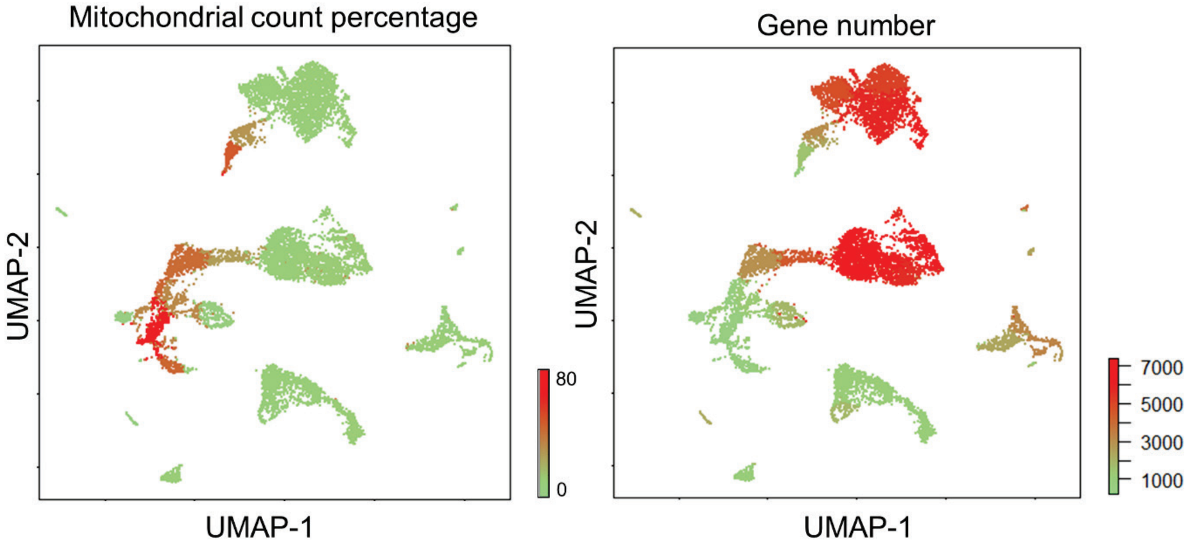


Figure S12. UMAP visualization of average mitochondrial count percentage (left) and gene number (right) of the cell clusters identified in Fig. 4A. Both of them are quality-control metrics commonly used to identify stressed and dying cells.

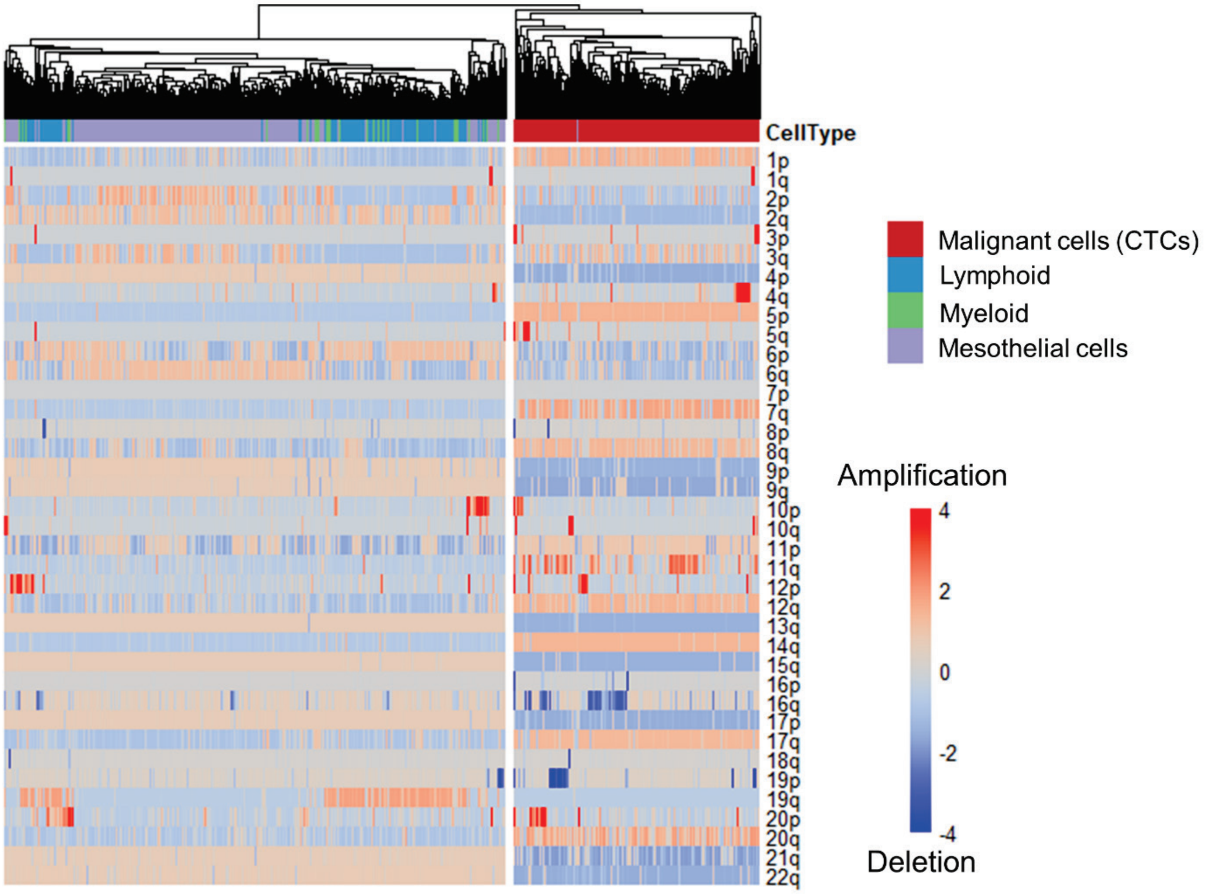


Figure S13. K-mean clustering of inferred single-cell CNV profiles across the chromosome. Chromosome arms that failed to fit CONICS two-component Gaussian mixture model were excluded. Cell types were labeled based on the UMAP clustering results in Fig. 4A. All the malignant cells (CTCs) are clustered to the right with significantly altered CNV profiles across the chromosome compared to non-malignant cells in the left cluster. Stressed/dying and erythroid cells are excluded from the analysis.

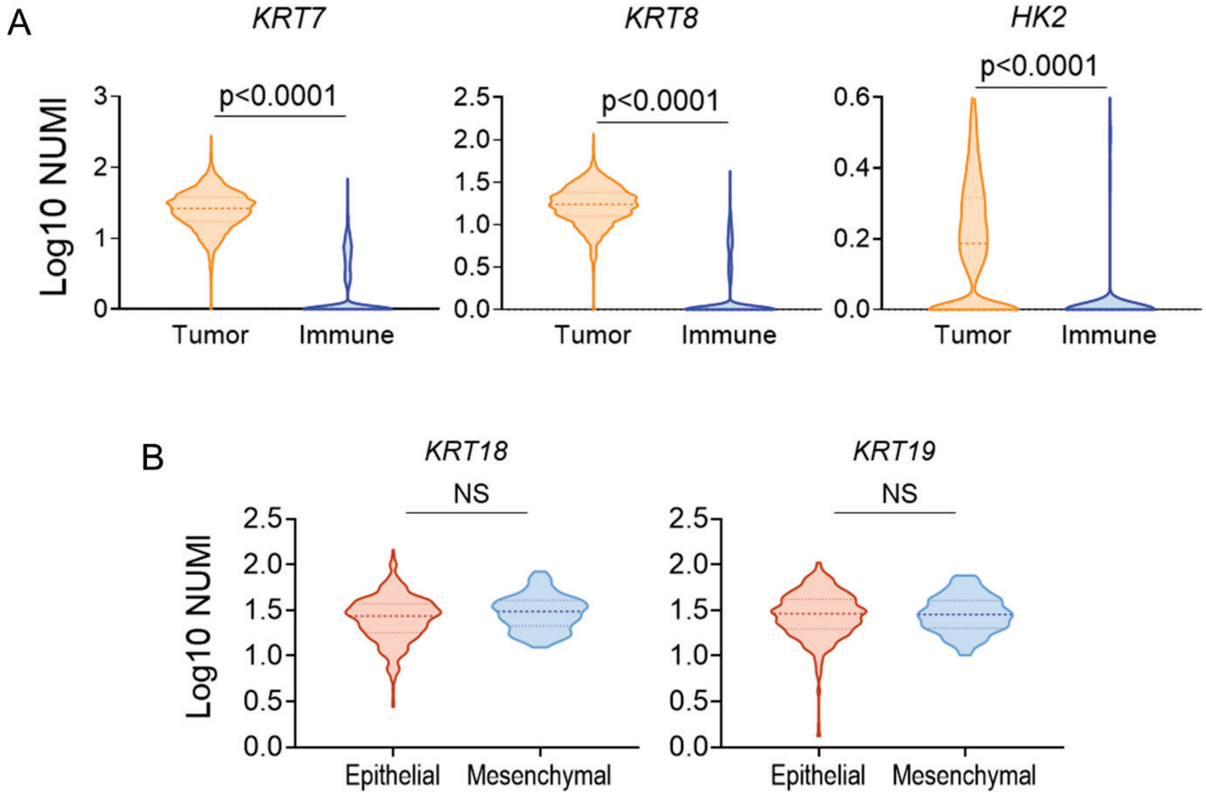


Figure S14. (A) Comparison of *KRT7*, *KRT8,* and *HK2* expression levels between CTCs and immune cells (N=2424 and 2009, respectively) from the single-cell transcriptome sequencing of the MPE sample. The dashed and dotted lines of each violin plot denote the median and first and third quartiles, respectively. (B) Comparison of expression levels of *KRT18* and *KRT19* between epithelial and mesenchymal CTCs (N=426 and 58, respectively, NS: not significant). The dashed and dotted lines of each violin plot denote the median and first and third quartiles, respectively.

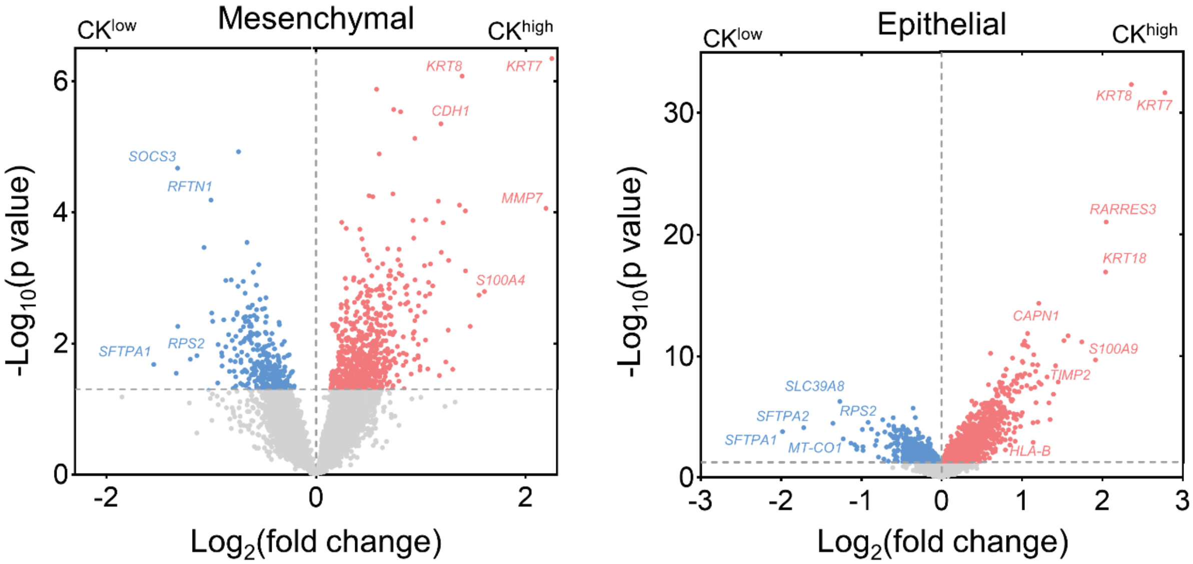


Figure S15. Volcano plots showing the DEGs between CK^{high} and CK^{low} CTCs in the epithelial and mesenchymal populations identified from the single-cell transcriptome data of the MPE sample. Statistically insignificant DEGs (p>0.05) are colored in grey.

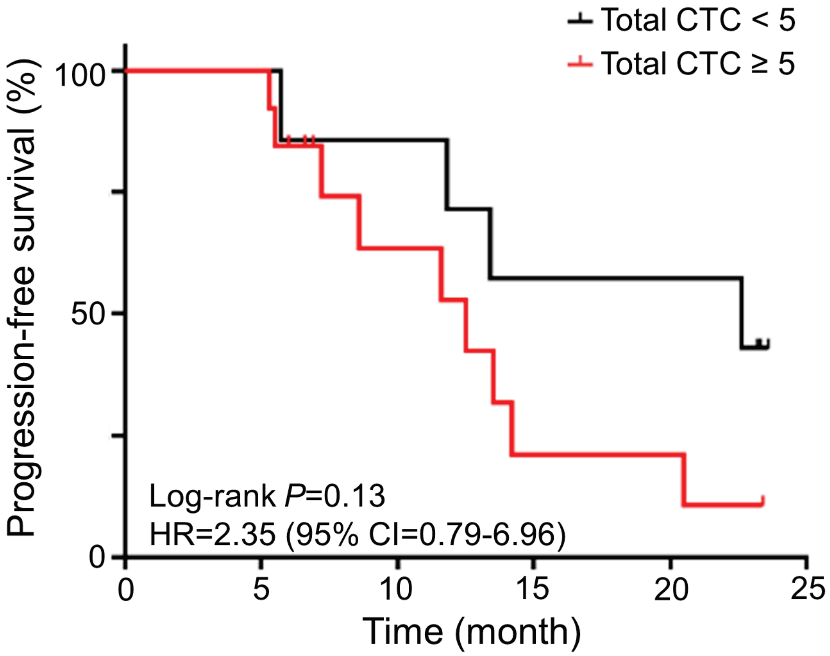
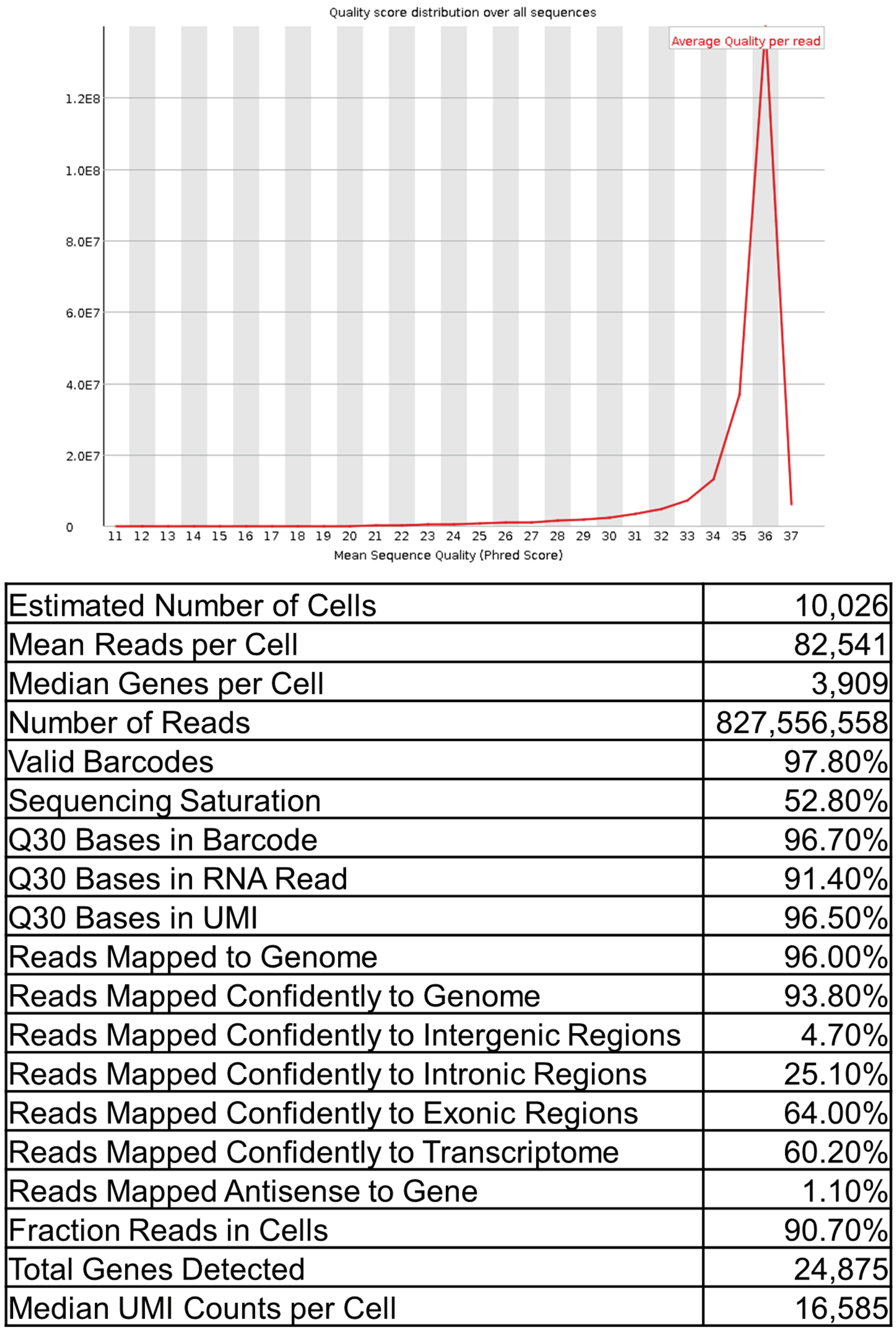


Figure S16. Kaplan-Meier curves showing PFS of the treatment naïve LUAD patients with positive CTC counts in blood segregated by the number of total CTCs (patient N=20). The log-rank p-value and hazard ratio are indicated.



Estimated Number of Cells	10,026
Mean Reads per Cell	82,541
Median Genes per Cell	3,909
Number of Reads	827,556,558
Valid Barcodes	97.80%
Sequencing Saturation	52.80%
Q30 Bases in Barcode	96.70%
Q30 Bases in RNA Read	91.40%
Q30 Bases in UMI	96.50%
Reads Mapped to Genome	96.00%
Reads Mapped Confidently to Genome	93.80%
Reads Mapped Confidently to Intergenic Regions	4.70%
Reads Mapped Confidently to Intronic Regions	25.10%
Reads Mapped Confidently to Exonic Regions	64.00%
Reads Mapped Confidently to Transcriptome	60.20%
Reads Mapped Antisense to Gene	1.10%
Fraction Reads in Cells	90.70%
Total Genes Detected	24,875
Median UMI Counts per Cell	16,585

Figure S17. Quality controls for the single-cell RNA sequencing via 10X Genomics. Top, quality score distribution showing that most of the sequencing reads are higher than Q30. Bottom, various quality control metrics extracted from Cell Ranger pipelines. These metrics confirmed that the sequencing quality of the sample meets regular standards of single-cell RNA-seq data generated by 10X Genomics.

III. Supplementary Tables

Table S1. Clinical information and CTC measurement results of blood samples from Stage III/V treatment naïve LUAD patients in this study. P19 was excluded due to stage II of LUAD. P20 and P23 were excluded because of prior treatment.

No	Age	Sex	Targetable Mutation	Stage	Metastatic sites	Number of CTC subtypes			No. of total CTC	PFS (month)	Immunohistochemistry of tumor tissues
						$HK2^{high}/CK^{neg}$	$HK2^{low}/CK^{pos}$	$HK2^{high}/CK^{pos}$			
1	72	F	$EGFR^{L858R}$	IV	Lung, bone	51	1	18	70	11.6	TTF-1+/CK+/NapsinA+/P40-/CD56-
2	68	F	$EGFR^{19Del}$	IV	Lung, liver, bone, adrenal gland	1	117	2	120	5.5	TTF-1+/CK+/NapsinA+/P40-/CD56-**
3	64	M	None	IV	Lung	10	0	0	10	8.6	TTF-1+/CK+/NapsinA+/P40-/CD56-
4	61	F	$EGFR^{L858R}$	IV	Brain, bone	8	3	0	11	14.2	TTF-1+/CK+/NapsinA+/P40-/CD56-
5	72	F	$EGFR^{L858R}$	IV	Lung, adrenal gland	5	0	0	5	7.2	TTF-1+/CK+/NapsinA+/P40-/CD56-
6	48	F	$EGFR^{L858R}$	IV	Pleura, bone, liver	3	0	0	3	11.8	TTF-1+/CK+/NapsinA+/P40-/CD56-**
7	67	F	$EGFR^{19Del}$	IV	Lung, pleura, bone	0	2	0	2	13.4	TTF-1+/CK+/NapsinA+/P40-/CD56-
8	50	F	$EGFR^{L858R}$	IV	Bone	0	0	5	5	20.5	TTF-1+/CK+/NapsinA+/P40-/CD56-
9	55	M	None	IV	Pleura	7	10	0	17	13.5	TTF-1+/CK+/NapsinA+/P40-/CD56-
10	60	F	None	IV	Bone	0	3	0	3	23.3*	TTF-1+/CK+/NapsinA+/P40-/CD56-
11	70	M	$EGFR^{L858R}$	IV	Pleura, liver	2	0	0	2	5.7	TTF-1+/CK+/NapsinA+/P40-/CD56-
12	66	F	$EGFR^{19Del}$	IIIC	None	3	0	0	3	22.6	TTF-1+/CK+/NapsinA+/P40-/CD56-
13	38	F	$EGFR^{19Del}$	IIIB	None	1	0	1	2	23.6*	TTF-1+/CK+/NapsinA+/P40-/CD56-
14	37	F	$EGFR^{19Del}$	IV	Lung	0	1	0	1	23.2*	TTF-1+/CK+/NapsinA+/P40-/CD56-
15	61	M	$EGFR^{19Del}$	IIIB	None	3	0	3	6	23.4*	TTF-1+/CK+/NapsinA+/P40-/CD56-
16	66	M	$EGFR^{L858R}$	IIIB	None	8	0	0	8	12.5	TTF-1+/CK+/NapsinA+/P40-/CD56-
17	63	F	$EGFR^{L858R}$	IV	Bone	0	0	0	0	-	TTF-1+/CK+/NapsinA+/P40-/CD56-
18	54	F	$EGFR^{19Del}$	IV	Lung	0	0	0	0	-	TTF-1-/CK+/NapsinA-/P40-/CD56-/CK7+
21	68	M	EGFR^{19Del}, EGFR^{G719X}	IV	Brain, adrenal gland	0	0	0	0	-	TTF-1-/CK+/NapsinA-/P40-/CD56-/CK7+
22	56	M	None	IV	Brain	0	0	0	0	-	TTF-1-/CK+/NapsinA+/P40-/CD56-

24	68	M	*EGFR*^{G719X}, *EGFR*^{L861Q}	IIIB	None	0	0	0	0	-	TTF-1+/CK+/NapsinA+/P40-/CD56-
34	55	M	None	IV	Brain, pleura	8	0	2	10	5.3	TTF-1+/CK+/NapsinA+/P40-/CD56-
35	55	F	None	IIIB	None	2	7	0	9	6.9*	TTF-1+/CK+/NapsinA-/P40-/CD56-
36	64	M	*EGFR*^{L858R}	IV	Lung, pleura, bone	8	0	0	8	6.6*	TTF-1+/CK+/NapsinA+/P40-/CD56-
37	69	F	*EGFR*^{19Del}	IV	Bone	0	13	2	15	6.0*	TTF-1+/CK+/NapsinA+/P40-/CD56-
38	66	M	*EGFR*^{19Del}	IIIA	None	0	0	0	0	-	TTF-1+/CK+/NapsinA+/P40-/CD56-
39	77	M	None	IIIA	None	0	0	0	0	-	TTF-1+/CK+/NapsinA+/P40-/CD56-
40	61	F	*EGFR*^{19Del}	IV	Brain, bone	3	1	0	4		TTF-1+/CK+/NapsinA+/P40-/CD56-
41	61	M	*EGFR*^{19Del}	IV	Liver,bone, pleura	14	10	5	29		TTF-1+/CK+/NapsinA+/P40-/CD56-
42	77	F	*EGFR*^{19Del}	IIIA	None	2	0	0	2		TTF-1+/CK+/NapsinA+/P40-/CD56-
43	41	M	*EGFR*^{19Del}	IV	Bone	2	36	1	39		TTF-1+/CK+/NapsinA+/P40-/CD56-
44	72	M	None	IV	Lung, pleura	0	11	0	11		TTF-1+/CK+/NapsinA+/P40-/CD56-
45	64	M	*EGFR*^{19Del}	IV	Bone, brain	4	5	6	15		TTF-1+/CK+/NapsinA+/P40-/CD56-
46	65	M	None	IIIC	None	1	1	0	2		TTF-1+/CK+/NapsinA+/P40-/CD56+
47	77	M	None	IIIB	None	3	17	3	23		TTF-1+/CK+/NapsinA+/P40-/CD56+
48	64	M	*EGFR*^{L858R}	IV	Bone	7	2	0	9		TTF-1+/CK+/NapsinA+/P40-/CD56+
49	65	M	*EGFR*^{L858R}	IIIB	None	1	0	0	1		TTF-1+/CK+/NapsinA+/P40-/CD56-
50	61	M	None	IIIB	None	3	0	0	3		TTF-1+/CK+/NapsinA+/P40-/CD56+
51	52	F	None (*ALK* fusion)	IV	Lung, pleura, bone	1	4	0	5		TTF-1+/CK+/NapsinA+/P40-/CD56-

52	55	F	$EGFR^{L858R}$	IIIA	None	0	1	0	1		TTF-1+/CK+/NapsinA+/P40-/CD56+
53	65	F	$EGFR^{L858R}$	IV	Lung, pleura, bone, brain	10	0	0	10		TTF-1+/CK+/NapsinA+/P40-/CD56- **
54	36	F	$EGFR^{L858R}$	IIIB	None	3	1	0	4		TTF-1+/CK+/NapsinA+/P40-/CD56-
55	65	M	None	IV	adrenal gland	3	7	0	10		TTF-1+/CK+/NapsinA-/P40-/CD56-
56	70	M	None	IIIB	None	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56+
57	56	M	$EGFR^{L858R}$	IV	Liver, bone	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56-
58	45	M	$EGFR^{19Del}$	IV	Pleura, liver	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56-
59	71	M	None	IV	Bone	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56+
60	54	F	$EGFR^{L858R}$	IV	Brain, bone	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56-
61	77	M	None	IIIA	None	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56+
62	72	M	None	IIIB	None	0	0	0	0		TTF-1+/CK+/NapsinA+/P40-/CD56-

*Not defined; **results from cell blocks of malignant pleural effusions

Table S2. Clinicopathological characteristics and CTC counts of treatment naïve LUAD patients with blood draws in this study. IQR, interquartile range.

Variable	**Evaluable Patient (N=50)**
Age, median (Range)	64 (36-77)
Sex (%)	
Female	22 (44%)
Male	28 (56%)
Tumor stage	
III (%)	18 (36%)
IV (%)	32 (64%)
Targetable mutation (%)	
EGFR	32 (64%)
ALK	1 (2%)
None	17 (34%)
No. of site of metastasis (%)	
0	18 (36%)
1	13 (26%)
2	12 (24%)
3+	7 (14%)
CTC count, median (IQR)	
Stage III patient	2 (0-3.75)
Stage IV patient	5 (0.75-11)

Table S3. Comparison of CTC detection in patients with NSCLC between EpCAM/CK-based assays and the HK2-based assay

Group (reference)	Patient disease stage	Size of cohort	Baseline positivity rate (%)	Definition of positivity
Krebs et al. (1)	III–IV	101	21	>=2 CTCs in 7.5mL blood
Hofman et al. (2)	I–IV	210	39	>=1 CTCs in 7.5mL blood
	III-IV	79	39	>=1 CTCs in 7.5mL blood
Tanaka et al. (3)	I–IV	110	24	>=1 CTCs in 7.5mL blood
	III-V	32	38	>=1 CTCs in 7.5mL blood
Isobe et al. (4)	IV	24	33	>=1 CTCs in 7.5mL blood
			17	>=2 CTCs in 7.5mL blood
Hirose et al. (5)	IV	33	36	>=1 CTCs in 7.5mL blood
Marchetti et al. (6)	EGFR mutation	37	41	>=1 CTCs in 7.5mL blood
Devriese et al. (7)	IIIB–IV	46	30	>=1 CTCs in 7.5mL blood
Lindsay et al. (8)	IIIB–IV	125	54.4	>=1 CTCs in 7.5mL blood
			40.8	>=2 CTCs in 7.5mL blood
Our HK2-based assay	**III-IV**	**50**	**72**	**>=1 CTCs in 5mL blood**
			66	**>=2 CTCs in 5mL blood**

(1) Krebs MG, et al. Evaluation and prognostic significance of circulating tumor cells in patients with non-small-cell lung cancer. *J Clin Oncol*. (2011) 29(12):1556-63.

(2) Hofman V, et al. Detection of circulating tumor cells as a prognostic factor in patients undergoing radical surgery for non-small-cell lung carcinoma: comparison of the efficacy of the CellSearch Assay™ and the isolation by size of epithelial tumor cell method. *Int J Cancer* (2011) 129(7):1651-60.

(3) Tanaka F, Yoneda K, Kondo N, Hashimoto M, Takuwa T, Matsumoto S, et al. Circulating tumor cell as a diagnostic marker in primary lung cancer. *Clin Cancer Res* (2009) 15:6980–6.10.1158/1078-0432.CCR-09-1095

(4) Isobe K, Hata Y, Kobayashi K, Hirota N, Sato K, Sano G, et al. Clinical significance of circulating tumor cells and free DNA in non-small cell lung cancer. *Anticancer Res* (2012) 32:3339–44

(5) Hirose T, et al. Relationship of circulating tumor cells to the effectiveness of cytotoxic chemotherapy in patients with metastatic non-small-cell lung cancer. *Oncol Res*. (2012) 20(2-3):131-7

(6) Marchetti A, et al. Assessment of EGFR mutations in circulating tumor cell preparations from NSCLC patients by next generation sequencing: toward a real-time liquid biopsy for treatment. *PLoS One* (2014) 9(8):e103883

(7) Devriese LA, et al. Circulating tumor cell detection in advanced non-small cell lung cancer patients by multi-marker QPCR analysis. *Lung Cancer* (2012) 75(2):242-7

(8) Lindsay CR, et al. A prospective examination of circulating tumor cell profiles in non-small-cell lung cancer molecular subgroups. *Ann Oncol* (2017) 28(7):1523-1531.

Table S4. Clinical information and CTC measurement results of blood samples from 30 healthy donors

Donor No.	Age	Sex	No. of CTC
1	33	M	0
2	72	M	0
3	41	F	0
4	48	F	0
5	63	F	0
6	70	M	0
7	51	M	0
8	48	F	0
9	40	M	0
10	32	F	0
11	34	M	0
12	44	F	0
13	63	M	0
14	58	M	0
15	61	M	0
16	49	F	0
17	55	F	0
18	39	M	0
19	75	M	0
20	60	M	0
21	55	M	0
22	58	F	0
23	43	F	0
24	65	M	0
25	69	M	0
26	49	F	0
27	52	M	0
28	35	F	0
29	43	F	0
30	67	M	0

Table S5. Clinical information and CTC measurement results of MPE samples from treatment-naïve LUAD patients in this study.

No	Age	Sex	Stage	Targetable mutation	Volume (mL)	Number of CTC subtypes		CK^{neg} CTC (%)
						$HK2^{high}/CK^{neg}$	$HK2^{high}/CK^{pos}$	
25	70	M	IV	EGFR^{L858R}	5	25	100	20

26	75	M	IV	*EGFR*^{19Del}	10	12	355	3.3
27	84	F	IV	*EGFR*^{19Del}	5	32	3939	0.8
28	73	M	IV	None	0.7	18	11760	0.2
29	67	F	IV	None	5	32	3040	1.1
30	65	F	IV	*EGFR*^{L858R}	1.5	9	260	3.5

Table S6. Clinical information and CTC measurement results of CSF samples from treatment-naïve LUAD patients in this study.

No	Age	Sex	Stage	Targetable mutation	Volume (mL)	Number of CTC subtypes			CK^{neg} CTC (%)
						HK2^{high}/CK^{neg}	HK2^{high}/CK^{pos}	HK2^{low}/CK^{pos}	
31	52	M	IV	*EGFR*^{L858R}	1.5	8	258	1564	0.44
32	37	F	IV	*EGFR*^{19Del}	1.5	11	202	77	3.8
33	51	F	IV	*EGFR*^{L858R}	1.5	3	11	114	2.3

Table S7. Primers used in this study. All primers were synthesized by Genewiz.

Primer Name	Sequence (5'-3')
EGFR-exon19-F	GTGGCACCATCTCACAATT
EGFR-exon19-R	ATGCTCCAGGCTCACCAAG
EGFR-exon21-F	TTCGCCAGCCATAAGTCCT
EGFR-exon21-R	TCATTCACTGTCCCAGCAAG
Chr1F	TTTAGGCGTCATCTGAGGGTA
Chr1R	TGGCAGCAGTATGGAGAATGTA
Chr2F	AGCGGGAGGGACTATTTCAC
Chr2R	GGATCGTTCAAAGGGAACTG
Chr3F	CCCTTGTACTGGCTCGTGTT
Chr3R	CTTGCACATGAAGGTCTGGA
Chr4F	GAGCATCTCTTGGCTCTGCT
Chr4R	TTGGGAAAGCACAGATCCTT
Chr5F	ACGGACAGTGGACAGATTGC
Chr5R	CCACTGTGCCACCCCATT
Chr6F	GAGGAGGGCAAGGAGAGAGT
Chr6R	ACCCTCCAGTGTGCAAAAAC
Chr7F	CTTCCTGCCATTCCACAAGT
Chr7R	CCCACTTTCATGCCTCTGAT
Chr8F	CTTCCCTGCCTTGCTCTCTA
Chr8R	CGGGACATTTCAGCAATCTT
Chr9F	CTGTGGAGCAGCTGTTTCTG
Chr9R	GAATTCACAAAGCCCCAAGA

Chr10F	CCCCTCATTCAAATCAGCAT
Chr10R	CAGGCAAAAGCTGGAGTTTC
Chr11F	TGAATGAGAACGCAGATGTGA
Chr11R	CACAAAGCATCCAGGGTCATT
Chr12F	ATCATGGAAATGCAGCCTCT
Chr12R	AGAACCCAGCTGGAATGATG
Chr13F	TGTTTCATGGAGTCCTGCTG
Chr13R	GGAGGCAAGAACCAAACAAA
Chr14F	AGCCAAGACGTACCCTCTCA
Chr14R	TGCTTTACACCAATCCCACA
Chr15F	TCAGCATGGGTTATGGGTTT
Chr15R	CCCAGATGATGGAGAGGAAA
Chr16F	GCCTGTGTTTGCTGATGAAA
Chr16R	GGGCAACGACCGTACTTAAA
Chr17F	TCCTGGGCTAGCCTTTTACA
Chr17R	ATCGCTTGAGCACTGAAGGT
Chr18F	AGACGAGCCTTTCTCTGTCG
Chr18R	TCGAGACCATCCCCACTAAC
Chr19F	AGTTGAGGAGATGGTGGAGC
Chr19R	AACAGGAGCCTTGGTCAGTC
Chr20F	CTGGTCAAACATCTCCCTCGT
Chr20R	CTCCACGCATCTTACATCACCT
Chr21F	GGACTTTGCTGACGGGATTA
Chr21R	GAACTAACGACCTCACGCTTG
Chr22F	TCCACAACCCCTTATCTTACCC
Chr22R	ACCTCAGGTGATCTACCCGC

IV. Supplementary Datasets

Dataset S1. Expression of the 200 EMT-defining genes and EMT enrichment scores of all the CTCs and TCGA patient samples.

Dataset S2. List of top 100 DEGs upregulated in CK^{high} (or CK^{low}) cells in epithelial and mesenchymal CTC populations.