의학박사 학위논문

Impact of intratumoral heterogeneity on drug responses : Unmasked by single-cell transcriptome analysis

단일 세포 전사체 분석을 통한 종양내 이질성에 따른 약물 반응성 예측에 관한 연구

2015 년 8 월

서울대학교 대학원

의과학과 의과학전공 (생화학)

김 규 태

A thesis of the Degree of Doctor of Philosophy

단일 세포 전사체 분석을 통한 종양내 이질성에 따른 약물 반응성 예측에 관한 연구

Impact of intratumoral heterogeneity on drug responses : Unmasked by single-cell transcriptome analysis

August 2015

The Department of Biomedical Sciences,
Seoul National University
College of Medicine
Kyu-Tae Kim

Impact of intratumoral heterogeneity on drug responses : Unmasked by single-cell transcriptome analysis

by

Kyu-Tae Kim

A thesis submitted to the Department of Biomedical Sciences in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Medical Science at Seoul National University College of Medicine

June 2015

Approved by Thesis Committee:

Professor JONG-IL KIM **Chairman**

Professor Inhee Mook **Vice chairman**

Professor Wonshik Han

Professor Murim Choi

Professor WOONG-YANG PARK

학위논문 원문제공 서비스에 대한 동의서

본인의 학위논문에 대하여 서울대학교가 아래와 같이 학위논문 제공하는 것에 동의합니다.

1. 동의사항

① 본인의 논문을 보존이나 인터넷 등을 통한 온라인 서비스 목적으로 복제할 경우 저작물의 내용을 변경하지 않는 범위 내에서의 복제를 허용합니다.

② 본인의 논문을 디지털화하여 인터넷 등 정보통신망을 통한 논문의 일부 또는 전부의 복제.배포 및 전송 시 무료로 제공하는 것에 동의합니다.

2. 개인(저작자)의 의무

본 논문의 저작권을 타인에게 양도하거나 또는 출판을 허락하는 등 동의 내용을 변경하고자 할 때는 소속대학(원)에 공개의 유보 또는 해지를 즉시 통보하겠습니다.

3. 서울대학교의 의무

① 서울대학교는 본 논문을 외부에 제공할 경우 저작권 보호장치(DRM)를 사용하여야 합니다.

② 서울대학교는 본 논문에 대한 공개의 유보나 해지 신청 시 즉시 처리해야 합니다.

논문 제목: Impact of intratumoral heterogeneity on drug responses

: Unmasked by single-cell transcriptome analysis

학위구분: **석사 □ · 박사 ☑**
학 과: Department of Biomedical Sciences
학 번: 2012-30575
연 락 처: 010-6484-5850
저 작 자: 김규태 (인)

제 출 일: 2015 년 7 월 28 일

서울대학교총장 귀하

ABSTRACT

Introduction: Understanding distinct genomic signatures of a patient's cancer is required to design and predict accurate therapeutic responses. Current approaches regarding heterogeneous cancer cells en masse as a pooled population, however, hardly reflect complete genomic landscape of tumor diversity, missing potentially important minor subclones' implications such as metastasis and drug resistance.

Methods: To dissect intratumoral heterogeneity and discover unique patterns of subclonal behaviors against drug treatment responses in functional modalities and signaling pathways out of heterogeneous population derived from a cancer patient, tumor transcriptome was characterized at single-cell resolution by utilizing single-cell RNA sequencing (scRNA-seq).

Results: First, in a lung adenocarcinoma PDX model*, individual cells showed mosaic expression of SNVs and differential gene expression. We could cluster the PDX cells into three distinct groups according to the presence of KRAS G12D mutation and transcriptome-based risk scores (RS). A single cell group with KRAS G12D/high RS was activated in the RAS-MAPK signaling pathway, and targeted by anti-cancer drugs such as docetaxel and BKM120. In comparison, a KRAS G12D/low RS group showed inactive RAS-MAPK signaling despite the expression of KRAS G12D. The drug-resistant population recapitulated gene expression signatures

of the KRAS G12D/ low-RS group.

Second, in a metastatic renal cell carcinoma PDX model, enriched subclonality was identified in a metastasis tumor with activated expression signatures of epithelial-mesenchymal transition and poor prognosis. Integrated analysis of transcriptome profiling and drug screening identified the most effective anti-cancer drugs. Furthermore, singularity of single cells with mutually exclusive activation of EGFR and SRC signaling pathways could suggest the potential of combination therapy, and its efficacy was validated in 2D and 3D *in vitro* and *in vivo* models.

Conclusions: Patient-derived xenograft cells showed heterogeneous profiles in SNVs and gene expression, which could cluster them into subclones with differential responses to anti-cancer drugs. Taken together, our approach of single-cell RNA sequencing on tumors provides insights into the more accurate strategy to identify minor but potentially important subclones that are relevant to drug resistance.

* This work is published in *Genome Biology* (1).

Keywords: Single cell analysis, Lung adenocarcinoma, Renal cell carcionma, Patient-derived xenograft, Tumor heterogeneity, Drug response

Student number: 2012-30575

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LUAD: lung adenocarcinoma

RCC: renal cell carcinoma

PDX: patient-derived xenograft

SNV: single-nucleotide variation

RS: risk score

RTK: receptor tyrosine kinase

WES: whole-exome sequencing

RNA-seq: RNA sequencing

PCA: principle component analysis

ITH: intratumoral heterogeneity

INTRODUCTION

1. Dissection of intratumoral heterogeneity (ITH)

Cancer is a composition of heterogeneous subpopulations that are originated from single transformed cells, and consequently becomes distinct of their genomic contents and phenotypic behaviors within a particular environment (2). Compared to inter-tumoral heterogeneity which refers to genetic diversity across cancer patients, intra-tumoral heterogeneity (ITH) may be regional due to the presence of subclones that dominate different parts of the tumor, or mosaiform which refers to cells with different properties that are closely intermingled within a patient.

By sequential acquisition of somatic mutations and malignant phenotypic conversion, cancer cells are likely to adapt to survive and thrive in the foreign microenvironments (3). Peter Nowell previously suggested a model of tumor progression as a branched rather than linear evolutionary pattern in 1976 (4), and this theory has been proved by recent efforts to characterize spatiotemporally heterogeneous tumors. The mutational spectrum changes within a patient have been investigated across multi-regional tumors (5-9) or longitudinal samples upon therapies (10-12). Dissecting such diverse cellular behaviors enables us to develop therapeutic paradigms that not only target specific drivers but also the evolutionary of these drivers.

2-1. Lung adenocarcinoma

The identification of specific mutations in cancer has led to the development of molecularly targeted therapy to improve the survival of patients (13-15).

Especially lung adenocarcinoma (LADC), the most common histological subtype of non-small cell lung cancer (NSCLC) (16), can be denoted by genetic alterations in the receptor tyrosine kinase/RAS/RAF pathway (14), and testing mutations in EGFR or ALK fusion have become routine clinical practice (17). While these common mutations are current standards for predicting targeted drug sensitivity in LADC, they only cover a fraction of patients. To successfully implement molecular targeted therapy, efforts are ongoing to catalogue different genomic signatures of different patients, i.e., inter-tumoral heterogeneity in genome-wide gene expression profiles, copy-number changes, and various types of mutations (13-15). These molecular signatures further foretell the clinical outcome of patients (18-21) and may in future increase the number of molecular targeted therapies; however, unexpected drug resistance and cancer progression even after successful targeted therapy (22) may occur and limit the clinical response.

2-2. Renal cell carcinoma

Clear cell renal cell carcinoma (ccRCC) initiated from the renal epithelium is the most prevalent histological type of adult kidney cancers. Dissecting intratumoral heterogeneity (ITH) of ccRCC has leveraged to extend our knowledge on how primary tumors harboring driver mutations evolve and spread to other sites (6, 23). The cellular fractions within and across the primary RCC (pRCC) and metastatic RCC (mRCC) are heterogeneous in both their genetic and biological features determining the variability in clinical

aggressiveness and sensitivity to the therapy (24-26). To achieve sustainable therapeutic benefit with targeted agents in mRCC, the effective target should focus on signaling pathways that are related to driver mutations occurred early in the clonal evolution of the disease and thus should be common to primary tumor and metastatic sites (27). Considering that extensive genetic heterogeneity may result in drug response variability among patients and treatment resistance, the tailored strategies for metastatic RCC is urgently needed.

3. Patient-derived Xenograft (PDX) model

A major reason why many preclinical findings based on relatively homogeneous established human cancer cell lines fail to translate directly into clinical applications include the lack of the cellular heterogeneity of the parental tumors (28). The transplantable, patient-derived cancer tissue xenografts (PDX) better represent the molecular, genetic, and histopathologic heterogeneity of the original tumors through serial passaging in mice (29, 30), which allows for capturing of the inter-tumoral and intra-tumoral heterogeneity in a wide spectrum of cancer types at the cellular level (31). Personalized tumorgraft models, also called “avatars”, propagated using patient-derived tumors have shown some success when used to guide clinical treatment in patients with advanced cancer (32, 33).

4. Single-cell transcriptome analysis

Single-cell sequencing technology has been rapidly progressed over the last few years with potentials to observe individual cellular characteristics that are usually masked in stochastic average of bulk cell level (34). Methodologies of isolating cells from sensitive and controllable devices (35-37) and amplifying initial contents of DNA (38, 39) or RNA (40, 41) in quasi-linear fashions have been improved to avoid getting skewed signals from single cells. Such previous efforts enable to accelerate the highest-resolution profiling of ITH.

Because RNA sequence mirrors the sequence of DNA from which it was transcribed, the entire collection of RNA sequences (transcriptome) in a cell allows us to understand unique behaviors of gene activity reflecting to cellular surviving strategy in the context of its microenvironment (3). Diverse cellular responses to drug treatment were observed at the single level (42). Profiling of transcriptome dynamics in a pathway-based approach could provide comprehension to predict response to targeted therapeutics (43). Therefore, correlating the genotype-phenotype relationship in genetically distinct single cells can provide new insights for selecting the most appropriate clinical intervention for targeting such profiled single cells.

In this study, transcriptome profiling was performed on single PDX cells respectively from a LUAD (Part-I, Figure 1A) and a metastatic RCC (Part-II, Figure 1B) patient to elucidate the molecular mechanisms and underlying genomic characteristics relevant to anti-cancer drug treatments. Single-cell transcriptome analysis uncovered heterogeneous behaviors of individual tumor cells and provided new insights in drug resistance signatures that were

masked in bulk tumor analyses.

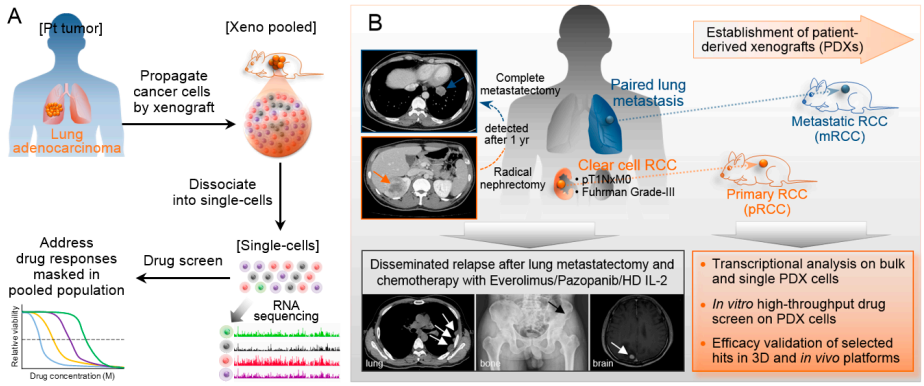


Figure 1. Experimental overview of this study. Schematic representations of Part-I (A) and Part-II (B). Parental tumors from a LUAD and a metastatic RCC patient were propagated by xenograft transplantation in humanized immunocompromised female NOG mice. The Part-I is designed to identify tumor cell subgroups associated with anti-cancer drug resistance with a LUAD PDX case. And, the Part-II shows how single-cell expression signatures could be translated into effective therapeutics with a metastatic RCC case.

MATERIALS AND METHODS

1. Patient samples and patient-derived xenograft (PDX) cells

This study was carried out in accordance with the principles of the Declaration of Helsinki, and approved by The Samsung Medical Center (Seoul, Korea) Institutional Review Board (IRB) (No. 2010-04-004). Surgical specimens were acquired from a 60-year-old male patient who underwent surgical resection of a 37-mm-sized irregular primary lung lesion in the right middle lobe (LC-PT-45), from a 57-year-old female patient who underwent surgical resection of a metachronous brain metastasis (LC-MBT-15), and from a 43-year old male who underwent radical nephrectomy (pT1Nx; Fuhrman Grade 3, RCC-12-085T) and metastatectomy (RCC-12-085T-LM). LC-PT-45 and RCC-paired tumors were taken in a treatment-naïve status whereas LC-MBT-15 tumor was taken after standard chemotherapy and erlotinib treatments. Pathologic examination of the primary tumors revealed a poorly differentiated lung adenocarcinoma based on the World Health Organization criteria (44). The PDX cells were isolated and cultured *in vitro* as described previously (45-47). Briefly, surgically removed tumor tissues were directly injected into the subrenal space of 6-8 week-old humanized immunocompromised female NOG (NOD/Shi- SCID/IL-2Rγ-null) mice (Orient Bio, Seongnam, Korea). Xenograft tumors were taken from the mice for PDX cell culture and validated by short tandem repeats DNA fingerprinting as having been derived from the original tumor. We used PDX cells at fewer than 3 in vitro passages for single-cell RNA-seq and drug screening. Animal care and handling was performed according to the National Institute of Health Guide for the Care and Use of Laboratory Animals (NIH

publication No.80-23, revised 1978).

2. *in vitro* 2D Drug screening with PDX cells

Dissociated PDX cells were cultured in neurobasal media-A supplemented with N2 (×1/2, Life Technologies, Carlsbad, CA, USA), B27 (×1/2, GIBCO, San Diego, CA, USA), basic fibroblast growth factor (bFGF, 25 ng/mL, R&D Systems, Minneapolis, MN, USA), epidermal growth factor (EGF, 25 ng/mL, R&D Systems), neuregulin 1 (NRG, 10 ng/mL, R&D Systems), and insulin-like growth factor 1(IGF1, 100ng/mL, R&D Systems). The cells grown in these serum-free sphere culture conditions were seeded in 384-well plates (500 cells/well), and treated with a drug library (Selleck, Houston, TX, USA). The drug library was composed of targeted agents and cytotoxic chemotherapeutics, which were included in the clinical guideline or current clinical trial for the treatment of non-small cell lung cancer. After 3 days of incubation at 37°C in a 5% CO_2 humidified incubator, cell viability was analyzed using an adenosine triphosphate monitoring system based on firefly luciferase (ATPlite™ 1step; PerkinElmer, Waltham, CA, USA). Test concentrations for each drug were empirically determined to produce a clinically relevant spectrum of drug activity. Dose response curves and corresponding half maximal (50%) inhibitory concentration values (IC_{50}) were calculated using the S+ Chip Analyzer (Samsung Electro-Mechanics, Suwon, Korea) (48). For signal transduction assays under treatment with targeted agents, primarily cultured PDX cells were maintained overnight in

serum-free sphere culture condition without growth factors, incubated for 1 hour with each inhibitor, and pulsed with original culture medium supplemented for 15 minutes. For western blot analysis, cells were lysed in RIPA lysis buffer supplemented with 1× phosphatase inhibitors (PhosStop; Roche Diagnostics, Indianapolis, IN, USA) and a 1× protease inhibitor cocktail (Complete Mini; Roche Diagnostics). After centrifugation at 10,000 g for 5 minutes, the supernatant was harvested. Protein concentration was determined using a bicinchoninic acid protein assay kit (Thermo, Waltham, MA, USA). Equal amounts of protein were subjected to sodium dodecyl sulfate-polyacrylamide gels electrophoresis and transferred to polyvinylidene difluoride membranes (Whatman, Maidstone, UK), which were blocked in 5% skim milk or bovine serum albumin for 1 hour at room temperature, incubated with the indicated primary antibodies overnight, and then blotted with the appropriate secondary antibodies. Antibodies against phospho-EGFR (Tyr1068), phospho-SRC (Cell Signaling Technology, Beverly, MA, USA), EGFR, SRC and GAPDH (Santa Cruz Biotechnology, Dallas, TX, USA) were used.

3. Drug treatment in microfluidic drug screening device

Microfluidic drug screening device was made of PDMS (polydimethylsiloxane, Sylgard 184; Dow Corning, MI, USA) by conventional soft-lithography process with 200 microns high SU-8 (MicroChem, MA, USA) patterned silicon wafer. Fabricated device was

sterilized and bonded on a cover glass to enclose microchannels by an oxygen plasma treatment (Femto Science, Seoul, Korea). The device was then kept in oven at 80℃ for 24 hours to allow the surface to recover its hydrophobicity. The restored hydrophobicity of the microfluidic channel surface helps injected ECM form stable interface with the side channels.

Collagen type 1 (3.0 mg/ml, rat tail; Corning, NY, USA) was used as an ECM scaffold for the cells 3D embedded inside. Collagen solution was prepared in pH7 and of 2mg/ml concentration. It was diluted in a mixture of 10X phosphate buffered saline (PBS; Gibco, NY, USA) and sterilized deionized water. Its pH was adjusted by 0.5N NaOH. Dissociated cells were then suspended in the collagen solution at a density of 0.5 X 106 cells/ml. The suspension was injected into a center channel of the device and allowed to gel by incubating in 37°C and 5% CO2 for 30 minutes. Detail of the device preparation and gel filling procedure was described in previous reference (49). To avoid cells attaching to the microfluidic channel surface, the device was turn upside down every 5 minutes. After gel formation, media containing each drug candidate was filled into side channels. The media in channel was refreshed every 24 hours. Viability of cells was quantified in 4 and 7 days of culture with Live/Dead Viability Assay Kit (Molecular Probes Invitrogen, CA, USA) containing calcein AM and ethidium homodimner for identifying live (green) and dead (red) cells, respectively. Cells in the microfluidic device was placed in a 37°C and 5% CO2 incubator for 30 minutes, and then staining solution was refreshed by PBS. Cell viability was calculated as the number of live cells divided by total cell number. Number of cells was counted by

ImageJ software (Image Processing and Analysis in Java, NIH). Normalized viability of cells was acquired by dividing with viability of cells cultured in a pure medium condition.

4. *in vivo* Xenograft drug treatment

In vivo experiments were conducted in accordance with the Institute for Laboratory Animal Research Guide for the Care and Use of Laboratory Animals and following protocols approved by the IRB at the Samsung Medical Center (Seoul, Korea). Athymic nude mice were utilized (Orient Bio, Korea). Afatinib and dasatinib (Selleckchem, Houston, TX, USA) were stored as 50 mg/ml solutions dissolved as in DMSO at -80°C for use at indicated concentrations. The stored solutions were diluted in PBS containing 4% ethanol, 5% polyethylene glycol 400 and 5% Tween 80 for treatment. Athymic female nude mice (BALB/c-nu/nu), 6- to 8-week-old, were used to establish RCC12-085T-LM cell xenograft model for the drug intervention experiment. Briefly, RCC12-085T-LM cells (2 x 105) mixed 1:1 with Matrigel (BD Biosciences) were inoculated subcutaneously in the right flank of each mouse. Tumor diameters were measured with calipers twice per week and volume in mm calculated by the formula: tumor volume = (l x w2)/2, where l is the longest diameter of the tumor, w is the shortest diameter of the tumor. The mice were randomized into three groups (5 in each group) after the tumors reached a mean volume of about 100-150 mm3. The treatment groups included: 1) Afatinib group (every day, 20 mg/kg, orally) for up to

three weeks 2) Dasatinib group (every day, 30 mg/kg, orally) for up to three weeks 3) Control group, receiving oral administration of control vehicle for up to three weeks. Throughout the study, mice were weighed and tumors were measured with a caliper every 4 days. Tumor volume was calculated by the following formula: tumor volume = (l x w2)/2, where l is the longest diameter of the tumor, w is the shortest diameter of the tumor. Mean tumor volumes were calculated, and growth curves were established as a function of time. The error bars indicated the value of the standard error of the mean. When the tumors grew to proper size, the mice were euthanized and the tumors were excised.

5. Whole exome sequencing (WES) and data processing

Genomic DNA was extracted from PDX cells using the QIAamp® DNA Mini kit (Qiagen, Hilden, Germany) or QIAamp DNA Blood Maxi Kit (Qiagen). Exomes were captured using the SureSelect XT Human All Exon V5 kit (Agilent Technologies, Inc., Santa Clara, CA). The sequencing library was constructed and analyzed by the HiSeq 2000 or 2500 systems (Illumina, San Diego, CA, USA) using the 100-bp paired-end mode of the TruSeq Rapid PE Cluster kit and TruSeq Rapid SBS kit (Illumina). Mean target coverage for exome data was 153.4±26.99X.

Exome-sequencing reads were aligned to the hg19 reference genome using BWA-0.7.10 (50). Putative duplications were marked by Picard-1.93. Sites potentially harboring small insertions or deletions were realigned, and single-

nucleotide variants were called by applying GATK-3.2 (51) software with known variant sites identified from phase I of The 1000 Genomes Project and dbSNP-137 (52). To detect somatic mutations with increased sensitivity both in lower and higher allele frequencies (53), we used the caller programs of MuTect-1.1.5 (54) and VarScan2 (55).

Estimation of copy number variation (CNV) from WES was performed using the ExomeCNV software package (56) in default quantification mode. Circular binary segmentation was applied to determine the neighboring regions of DNA that exhibited a statistically significant difference in copy number. The output was also applied to infer tumor purity using AbsCNseq (57).

6. Array comparative genomic hybridization and data processing

Purified DNA from patient-derived tumor samples was labeled with Cy5-dUTP following the Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis protocol (Ver-7.3, Agilent). Together with reference DNA samples labeled with Cy3-dUTP, Cy5-labeled DNA was quantified for DNA concentration and each labeled-fluorescence intensity using ND-1000 Spectrophotometer (NanoDrop, DE, USA). Labeled test and reference samples were then hybridized to the SurePrint G3 Human CGH 4×180K Microarrays (Agilent), according to the manufacturer's standard protocol. The dual-colored fluorescence signals were scanned on the Agilent

Microarray Scanner (Agilent) and translated to log10 ratios using Feature Extraction software (Ver-11.0.1.1, Agilent).

From the CGH data, extracted signals were normalized to log2 ratios using the limma package (58). To detect significant breakpoints across thousands of probe-derived signals, we applied the circular binary segmentation (CBS) algorithm using the DNAcopy package (59). After smoothing the data to detect outliers within chromosomes 1-22, aberrant segments were determined applying the significance level of 1.0E-04 to accept change-points based on a maximum t-statistic. We classified the segmented results into copy losses when the log2 ratios were lower than -0.25, and copy gains when those were greater than 0.25. Considering sample-specific tumor purity and ploidy, somatic copy number alterations (SCNA) were adjusted by implementing the ABSOLUTE algorithm (60). To compare the patterns of SCNA across samples, segment values were averaged with 1kb binning along the chromosomes.

7. Isolation of single cells and RNA-seq

We used the C1™ Single-Cell Auto Prep System (Fluidigm, San Francisco, CA, USA) with the SMARTer kit (Clontech, Mountain View, CA, USA). For the original experiment, 44 cells were captured as a single isolate on a C1 chip (17-25 μm) as determined by microscopic examination, and 34 passed the required criteria for cDNA quantity and quality as measured with a Qubit® 2.0 Fluorometer (Life Technologies) and 2100 Bioanalyzer (Agilent). RNAs

from bulk cell samples were also amplified using a SMARTer kit with 10 ng of starting material. Libraries were generated using the Nextera XT DNA Sample Prep Kit (Illumina) and sequenced on the HiSeq 2500 using the 100-bp paired-end mode of the TruSeq Rapid PE Cluster kit and TruSeq Rapid SBS kit.

8. RNA-seq data processing

RNA-seq reads were aligned to the human genome reference (hg19) together with splice junction information of each sample using the 2-pass default mode of STAR_2.4.0d (61). Gene expression was quantified by implementing RSEM v1.2.18 (62) in default mode with Genecode 19 (63) annotation, and calculated as the sum of isoform expression. In Part-II, To identify not only differential but also common expression signatures between primary RCC and paired lung metastasis single cells to normal-featured signals, we employed expression profiles of normal kidney cortex from the GTEx portal (http://www.gtexportal.org; transcript read counts V3). Then, we normalized sample-to-sample variation by applying mean centroid.

Pre-processing steps for RNA-seq reads before calling variants were optimized by deduplication, splitting reads into exon segments, hard-clipping any sequences overhanging into the intronic regions, realigning reads and recalibration using GATK-3.2 (51). Then, variants were called in 'HaplotypeCaller' mode. Further filtering was applied to SNVs that were regarded to be potential false positives in RNA-seq by SNPiR (64). We

regarded only those SNVs which overlapped with WES as true positives. The overall process of calling and filtering the variants is summarized in Figure 2.

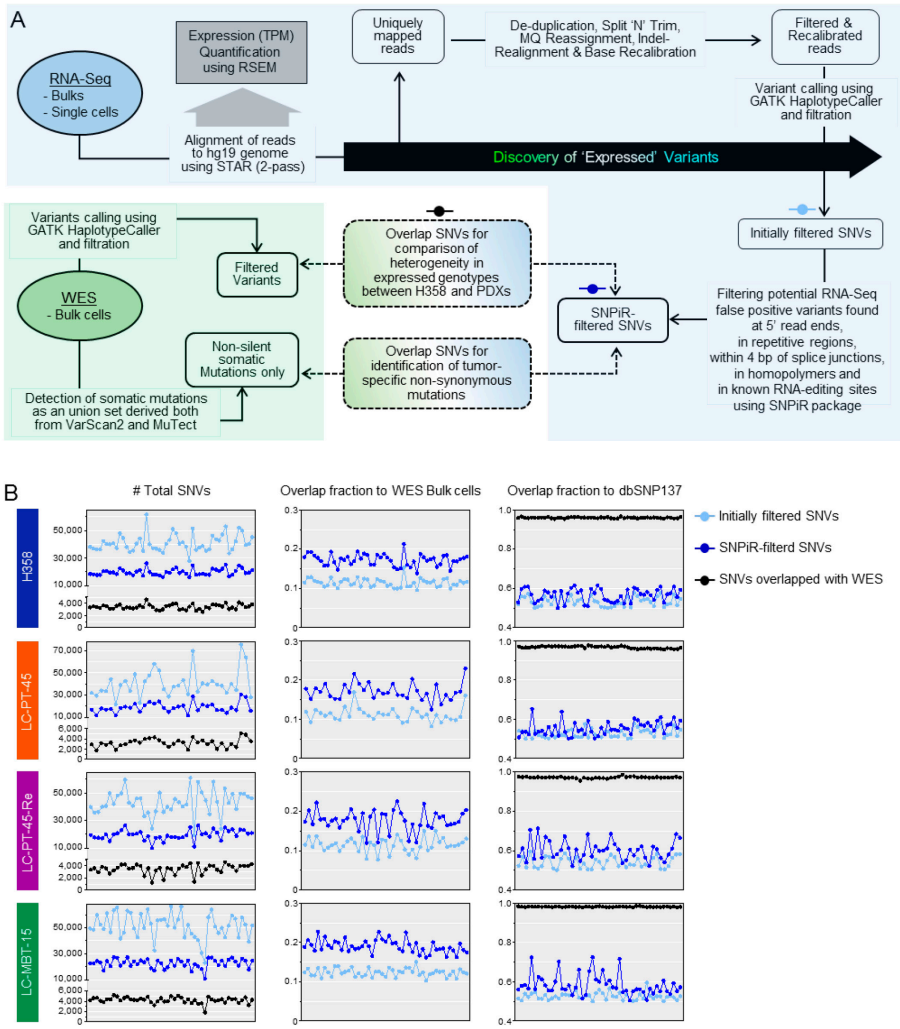


Figure 2. Detection and filtering of variants in single-cell RNA-seq data. (A) Schematic overview of data processing for the discovery of expressed variants. (B) Comparative evaluation of the detection processes for genomic variants in RNA-seq, following filtering steps marked in (A).

9. Clustering of genomic clones

To infer the subclonal structure between primary RCC and paired lung metastasis, we adopted the PyClone algorithm (65) that computes cellular prevalence of mutations and clusters these mutations based on a hierarchical Bayes statistical model. Mutational information including somatic SNVs called from the deep exomes and absolute copy-number changes corresponding to the region of SNVs was prepared to implement the PyClone. The cellular prevalence for each mutation was estimated using a beta-binomial model by setting the number of MCMC (Markov chain Monte Carlo) iterations to 100,000 with a burn-in of 50,000. The number of clusters was inferred from the average linkage hierarchical clustering in the post-burn-in trace by optimizing the MPEAR (maximization of posterior expected adjusted rand index) criterion.

10. Computing risk score using multivariate markers

Risk scores were regression coefficients calculated by a linear combination of the expression values of the prognosis markers using a training set (18) of LUAD patients. Prognosis markers were also derived from the previous report (18) that classified LUAD patients according to gene expression profiles of the suggested markers, and 69 genes were ultimately chosen by overlapping our data sets after gene filtering of zero expression across all single cells. These filtered genes (Table 2) were validated as prognosis markers with an independent dataset from TCGA LUAD (14). Batch effects on gene

expression between independent datasets were removed by means of R package ComBat. Regression coefficients and P-values of the training set were estimated using univariate Cox proportional hazards regression modeling and ordered by P-values. To partition patient samples into a high- and a low-RS based groups upon computed response score, we applied a 60th percentile cutoff as described in Beer D.G. *et al* (18). Survival analysis was performed using the R Survival package and validated through Kaplan-Meier survival curves with log-rank testing (training set, $P=1.04\times10^{-6}$; validation set, $P=9.25\times10^{-3}$).

To classify the control and drug-treated PDX cells into the semi-supervised clustered single cells (LC-PT-45, Fig. 4; LC-MBT-15, Fig. S11), the classification SVM Type 1 (C-SVM classification) model was applied using the R package e1071.

11. Gene set signature activation analysis

To characterize gene expression features of a subgroup compared to the other groups among the classified single cells, we utilized the GSEA-P program with default mode searching for significantly enriched gene set signatures. Applied gene sets were derived from the three major curated pathway databases of KEGG, REACTOME, and BIOCARTA in MSigDB v4.0. To estimate gene set activation status of single sample, Gene Set Variation Analysis (GSVA) (66) was applied in default mode. As to estimate GSVA scores applied in Part-II for the given gene sets as follows;

Labeled gene set	Original source
Stromal signature	(extracted from the ESTIMATE package (67))
EMT induced signature	(Taube J.H. *et al*., 2010 PNAS (68))
Prognostic signature	(The Cancer Genome Atlas Research Network for ccRCC, 2013 Nature (69))
EGFR signaling	[Reactome] Signaling by constitutively active EGFR
SRC signaling	(Extracted from Gatza M.L. *et al*., 2010 PNAS (70))
mTOR signaling	[Reactome] mTOR signaling
VEGFR signaling	[PID] Signaling events mediated by VEGFR1 and VEGFR2
RAF signaling	[Reactome] RAF activation
MEK signaling	[Reactome] MEK activation
c-Met signaling	[PID] Signaling events mediated by c-Met
SCF-KIT signaling	[Reactome] Signaling by SCF-KIT
PI3K/AKT signaling	[Reactome] PI3K/AKT Signaling in Cancer
FGFR signaling	[Reactome] Signaling by FGFR
PDGFR signaling	[PID] PDGF receptor signaling network

12. Validating gene expression and expressed SNVs at RNA level by qPCR

Gene expression variation between RNA-seq and qPCR across single cells was verified by using Biomark HD (Fluidigm). To compare correlations between the two technical platforms for the selected 43 genes, mean fold change over median expression was calculated as in the previous study (71). Validation of expressed SNVs at the RNA level was also carried out using Biomark HD (Fluidigm). Primers were designed using D3™ software (Fluidigm), and sequences are available in Table 3 and Table 4.

13. Validating genomic variants at DNA level by ddPCR

PDX cells were labeled with 6-Carboxyfluorescein succinimidyl ester (Life Technologies) and sorted into single cells using a FACSAria™ III flow

cytometer (BD Biosciences, CA, USA). Wells with a single green fluorescence signal were manually inspected and selected for amplification of genomic DNAs with a GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Little Chalfont, UK). The mutant alleles were detected using ddPCR Supermix for Probes reagents (Bio-Rad, Hercules, CA) implemented under QX200 ddPCR system, following the manufacturer's protocols. The negative signal of droplets was normalized with a vehicle control, and the numbers of wild-type or mutation alleles in droplets were estimated in Poisson distribution. Variant allele frequency (VAF) was calculated by counting mutation alleles over the total number of detected alleles. We regarded genotypes of detected variants as homozygous when the VAF was higher than 90%. Sequences of the primers used in ddPCR are available in Table 5.

14. Statistical analysis of single-cell gene expression

Linear regression was applied to scatter plots of the averaged single-cells over the pooled-cell samples in Figure 8A with zero intercepts. The inter-correlation distribution between single-cells was calculated as a Pearson's correlation coefficient and plotted as a density plot with a kernel function fitting over the histograms (Fig. 8B). Multiple regression analysis estimated how many single cells hypothetically accounted for the pooled cell fraction. Single-cell samples were randomly chosen with the given number and the coefficient R^2 (Fig. 8C) and the overlap ratio (Fig. 8E) were determined 1000 times with permutation. The differences in normalized RS, gene expression,

and gene set activation score between single-cell subgroups were tested using two-tailed Student's *t*-tests.

RESULTS (PART-I)

Identification of tumor cell subgroups associated with anti-cancer drug resistance in a lung adenocarcinoma patient

This research was published

in *Genome Biology* (1) on June 2015.

1. Intratumoral genetic heterogeneity of LUAD PDX cells

Surgically removed LUAD tissue was propagated through xenograft engraftments in mice. Viable cancer cells were dissociated from the PDX tissue and primarily cultured *in vitro* (Figure 4A). Cultured PDX cells were genetically analyzed by RNA sequencing (RNA-seq) and whole-exome sequencing (WES). Although the tumor portion in the surgical sample represented approximately 40% of the excised tissue volume, multiple validated genomic analyses utilizing WES (56, 57) and expression profiles (67) indicated that human cancer cells were highly enriched (~100%) in the PDX cells (Figure 3A). Overall, copy number alterations and variant allele frequencies were increased in the PDX tumor, compared to the surgical specimen (Figure 3A and 3B). The full profiles of somatic mutations in the patient tumor and PDX cells are listed in Table 1.

Tumor cell-enriched PDX cells (LC-PT-45) (47) were further analyzed by single-cell RNA-seq using the Fluidigm C1™ autoprep system with SMART-seq (40). cDNAs from 34 individual PDX cells were successfully amplified. Using 100-bp paired-end sequencing, we obtained an average of 8.12 ± 2.34 million mapped reads from the captured cells. Overall 85.63% of reads mapped to the human reference genome, which was a lower percentage that is typical for unamplified conventional RNA-seq, but comparable to other single cell RNA-seq data (40, 72). We also sequenced 50 single H358 human lung cancer cells as cell line controls and obtained an 85.39% mapping rate. Noticeably skewed coverage at the 3' end of transcripts, which was inversely proportional to the expression level (Figure 5), was observed in the single-cell

RNA-seq data. The use of smaller initial RNA templates for amplification is known to increase this bias (40).

Despite the sequencing bias in amplified RNAs, average gene expression in single cells correlated well with expression in bulk cells, for both H358 and PDX cells (Figure 8A). The inter-correlation distribution of gene expression among the 34 individual PDX cells was wider than that in the fifty H358 cells (Figure 8B), indicating higher transcriptome heterogeneity. The level of transcriptome heterogeneity was also evaluated by multiple regression analysis of different sized pools (n=5, 15, 25, 34/35, 50; randomly selected by permutation ×1000) of single cell transcriptomes to the bulk sample (Figure 8C). The modeling demonstrated that five H358 or PDX individual cells represented >70% of the gene expression of the whole population, with PDX cells showing wider variations in adjusted R-square dependent on randomly selected sample of five cells. With averaging increased number of cells, the single cell data approximated the bulk up to 85%, suggesting that the single cell data is consistent with the bulk data (Figure 8C). We repeated the single cell isolation and RNA-seq using 43 additional PDX cells and obtained comparable results that were highly correlated with the first data set (Figure 6 and Figure 8, LC-PT-45 and LC-PT-45-Re). Comparisons of gene expression data for the 43 target genes between technical replicate RNA-seq sets (Figure 6G left) or between RNA-seq and qPCR analysis (Figure 6G right) also indicated a good correlation, comparable to that reported in a previous publication (71).

2. Single-cell heterogeneity of expressed single-nucleotide variants

To estimate tumor heterogeneity at the genetic mutation level, we identified expressed SNVs using the single cell RNA-seq and bulk WES data (Figure 2A). After removal of potential false positive SNVs specifically found in RNA-seq, using the SNPiR package (64), higher overlap ratios to bulk WES data were observed (Figure 2B middle panels). Selection of SNVs co-found in single cell RNA-seq and bulk WES significantly increased the overlap ratios to dbSNP137 (Figure 2B right-side panels). These filtered SNVs of individual PDX cells were more heterogeneously expressed than those of H358 cells in terms of the lower overlap ratios between single cells (Figure 8D). Moreover, the union of SNVs from five PDX cells (randomly selected by permutation ×1000) reflected only 49% of the expressed SNVs in the whole population, whereas those of five H358 cells rep resented 75% (Figure 8E). With increased number of single cells, the coverage was increased up to 70% and 90% for PDX cells (34 LC-PT-45 or 43 LC-PT-45-Re) and H358 cells, respectively.

After exclusion of germline variants by selecting only somatic SNVs from bulk WES data, expression of fifty tumor-specific non-synonymous SNVs were analyzed in individual PDX cells (Figure 2A). The 50 tumor-specific SNVs were heterogeneously expressed in the individual PDX cells (Figure 9A, LC-PT-45), in comparison to less variable expression of COSMIC enlisted (73) lung cancer mutations for individual H358 cells (Figure 7). We detected comparable mutation patterns and frequencies in the original and replicate PDX analyses (Figure 9A, LC-PT-45-Re), which showed >70% concordance

with the qPCR based genotyping analysis (Figure 9A, right panel; Figure 9C; Figure 10). Among the SNVs detected in PDX cells, *KRAS* (13, 14), *GAPVD1* (74), and *JMJD1C* (75) were functionally related to the RTK-RAS-MAPK signaling pathway. The hotspot $KRAS^{G12D}$ mutation was detected in 27 out of 34 single PDX cells (79.4%), or 33 out of 43 PDX replicates (76.7%). Some cells with discrepant KRAS mutation calls between RNA-seq and qPCR genotyping had low levels of KRAS transcripts, indicating that RNA based genotyping is highly dependent on the gene expression level. To compare the relationship of mutation rates between RNA and DNA, we further assessed 13 somatic mutations at DNA level in another 29 PDX single cells by droplet digital PCR. Interestingly, this single cell DNA analysis revealed more cells with heterozygous genotypes compared to the expressed genotypes found in the RNA (Figure 11A), that might be related to allele-biased expression, which may add to sequence level transcriptome heterogeneity. The mutation rates computed as variant allele frequencies for bulk or as cellular frequencies for single cells showed overall high level of concordance between various methods of genotyping including both RNA and DNA sources (Figure 11B). With respect to *KRAS* mutation, most PDX cells (25/29) were positive, with *KRAS* allele drop-outs occuring in the remaining four cells. Of note, copy number gains (Figure 9C) and a wide range of mutant/wild-type ratios in *KRAS* ddPCR (data not shown) suggest that the differential expression levels of mutant or wild-type alleles might have caused heterogeneity in the expressed genotype. Given the importance of oncogenic *KRAS* mutations, we defined two subpopulations in the PDX based on the expressed genotype; one

with dominant *KRAS*^{G12D} expression, and another without *KRAS*^{G12D} expression (*KRAS*$^{wild\ type\ (WT)}$ expression or no/low *KRAS* expression).

3. Identification of PDX cell subgroups

To further identify subclones with possible phenotypic implications in the PDX cells, we utilized the expression profiles of 69 genes related to the clinical prognosis of LUAD patients (Table 2) (18) as multivariate markers to compute a risk score (RS) (Figure 12A). Previous study (18) defined high-RS population as upper 40% of RS (normalized RS>0). The prognostic significance of the RS was validated in two independent public datasets (Figure 13). Moreover, a higher RS was significantly associated with the *KRAS* mutation in the LUAD patient population (18) (Figure 12B), which is consistent with previously observed correlation of the *KRAS* mutation with worse clinical outcomes (17, 76).

Interestingly, individual PDX cells were calculated to have a wide range of RS distribution (Figure 12A). Eighteen out of the 34 PDX cells or 21 out of 43 of the replicate samples were determined to be high-RS. We found that PDX cells with *KRAS*^{G12D} expression tend to have a higher RS (Figure 12B), which correlated well with those of LUAD patients in clinical studies [6]. Altogether, semi-supervised clustering based on the expression of the *KRAS* mutation and RS classified the PDX cells into three major groups: Group 1, no *KRAS*^{G12D} (*KRAS*^{WT} or no KRAS)/low RS (n=6); Group 2, *KRAS*^{G12D}/low

RS (n=17); and Group 3, *KRAS*^{G12D}/high RS (n=10), and a minor group of no *KRAS*^{G12D}/high RS (n=1) (Figure 12C and D).

The three major groups defined above, displayed characteristic gene expression profiles that likely reflect the different phenotypes among individual PDX cells. In particular, Group 3 had enhanced gene expression signatures related to the activation of the RAS-MAPK signaling pathway (77, 78) (Figure 12E-H), which correlated well with *KRAS* mutational status. Group 3 PDX cells also showed significantly higher cell cycle gene mRNA expression (Figure 14C) (79). In contrast, despite having the *KRAS* mutation, Group 2 cells had fewer activation characteristics of the RAS-MAPK signaling pathway (Figure 12H) and had reduced expression of cell cycle-related genes (Figure 14C).

The distinct gene expression signatures among the three groups were visualized by a principal component analysis (PCA) plot using genes exclusively expressed by each group, with a criterion of at least a two-fold change in transcripts per million (TPM) ratio with statistical significance (t-test $P<0.05$) (Figure 12D). Although Group 2 cells showed a lower RAS-MAPK signaling pathway activation status, they had significantly upregulated expression of ion channel transport pathway-related genes (Figure 14B) which has been implicated in the drug resistance mechanism (80).

4. Phenotypic interpretation of PDX cell subgroups

The results above indicated that, in the PDX cell population, there is a specific subgroup (Group 3) that is predicted to be more aggressive than the other groups. This subset is characterized by a high RS, *KRAS* mutation, RAS-MAPK signaling pathway activation, and cell cycle-related gene upregulation. To determine whether individual cells associate with tumor phenotypic aggressiveness such as drug resistance, we screened the *in vitro* sensitivity of the PDX cells against a panel of 25 anti-cancer agents used in non-small cell lung cancer treatment (Figure 16). The PDX cells were highly sensitive to a variety of drug treatments, including docetaxel, and molecular pathway targeting agents. Among the identified agents, we focused on the MEK1/2 inhibitor selumetinib, and the PI3K inhibitors BKM120 and BEZ235 (PI3K/mTOR), because of their potential clinical benefits (81, 82). Other cytotoxic drugs, *e.g.*, carboplatin, and the Notch inhibitor DAPT, did not show any effects (Figure 15A). Although docetaxel, BKM120, BEZ235, and selumetinib showed tumoricidal effects, some PDX cells survived through the 3 day of treatment with these drugs when utilized at their reported IC_{50}.

When evaluated as a bulk population, PDX cells manifested Group 3-like characteristics with high RS and $KRAS^{G12D}$. Ineffective treatments with carboplatin or DAPT did not alter these properties of the group (Figure 15B-G). However, those PDX cells that survived the docetaxel, BKM120, BEZ235, or selumetinib treatments showed Group 2-like gene expression signatures; low RS (Figure 15B), slight decrease in total KRAS expression (Figure 15C), down-regulation of gene expression signatures associated with KRAS overexpression (Figure 15D), preservation of the mutant $KRAS^{G12D}$ expression

(Figure 15E), and down-regulation of RAS-MAPK signaling pathway (Figure 15F). Moreover, upregulation of ion channel transport genes (Figure 14B) and downregulation of cell cycle-related genes (Figure 14C) were observed in these treatment groups. From these results, we hypothesized that the drug resistant population may have been a subpopulation that were cell-cycle quiescent and with possibly higher transporter activity for the anti-cancer drugs. The overall gene expression signature represented by PCA confirmed the Group 2 cell-like properties of the drug resistant PDX cells, in a SVM model (Figure 15G). Together, these results indicated that the Group 2-like characteristics persisted after effective anti-cancer drug treatments.

We further determined whether the group 2-like population conveyed the low risk gene expression signature after anti-cancer drug treatment with selumetinib (Figure 17A). Interestingly, the low RS of surviving PDX cells was gradually reverted to a high RS after drug removal (Figure 17B). The KRAS over-expression signature (Figure 17D) and MAPK pathway activation (Figure 17F) recovered as well. By contrast, the level of total KRAS expression (Figure 17C) and mutational status (Figure 17E) were not altered by drug removal. The possible mechanisms of the dynamic nature of these gene expression signatures, such as epigenetic regulation or recovery of heterogeneity by clonal proliferation, need to be further elucidated.

5. Validation of analytical procedures in an independent lung cancer PDX case

To validate our strategy of using single cell RNA-seq data for subgroup identification, we used an independent set of PDX cells derived from a lung cancer-brain metastasis (LC-MBT-15) (47). The LC-MBT-15 PDX harbors an insertional mutation in EGFR Exon20, a well-known driver mutation in LUAD confering resistance to reversible EGFR inhibitors (83, 84). Single cells from LC-MBT-15 had less heterogeneous transcriptome and SNV expression compared to the *KRAS* mutant PDX cells (Figure 17A-E), which might have been caused by extensive clonal selection during serial anti-cancer treatments before PDX establishment (See the patient description in the Materials and Methods). Nonetheless, the LC-MBT-15 single cells were still clustered into two subgroups by the risk score, similar to the original PDX case (Figure 17F). In contrast to the $KRAS^{G12D}$ mutation, the *EGFR* mutation was modestly detected and showed no preferential expression in the high RS group (Figure 17G and H).

Drug screening on LC-MBT-15 cells was performed using 28 lung cancer drugs (Figure 18). LC-MBT-15 cells were highly sensitive to the irreversible EGFR/HER2 inhibitor afatinib and the c-Met inhibitor tivantinib while resistant to the reversible EGFR inhibitor erlotinib. When gene expression profiles for the drug-resistant populations were analyzed 3 days later, PCA of the single cells and application of a SVM model for drug treated populations revealed that the drug-resistant populations shared the gene expression signature with the low risk score group. Interestingly, upregulation of ion channel transport genes was also noted in the drug resistant populations (Figure 17K) similar to the low risk group single cells. These results are

consistent with the original LC-PT-45 PDX case, and further support the observation that (1) single cell profiles of a population reveal cells with drug resistant signatures and (2) the drug-resistant population may come from a subset with higher transporter activity and low cell proliferation activity.

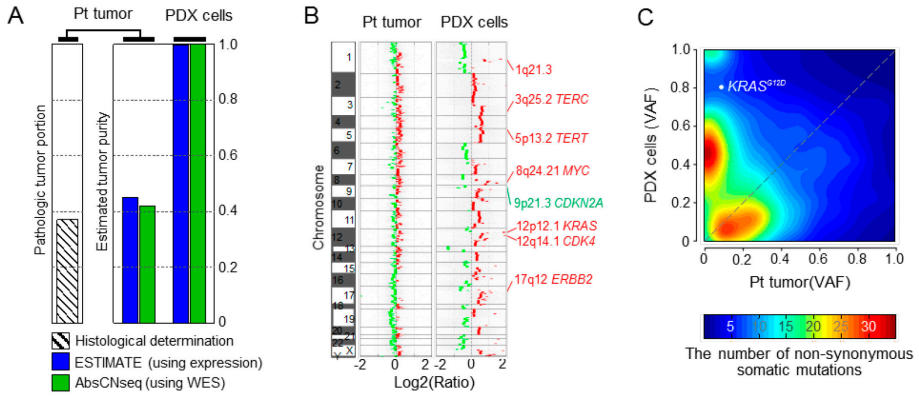
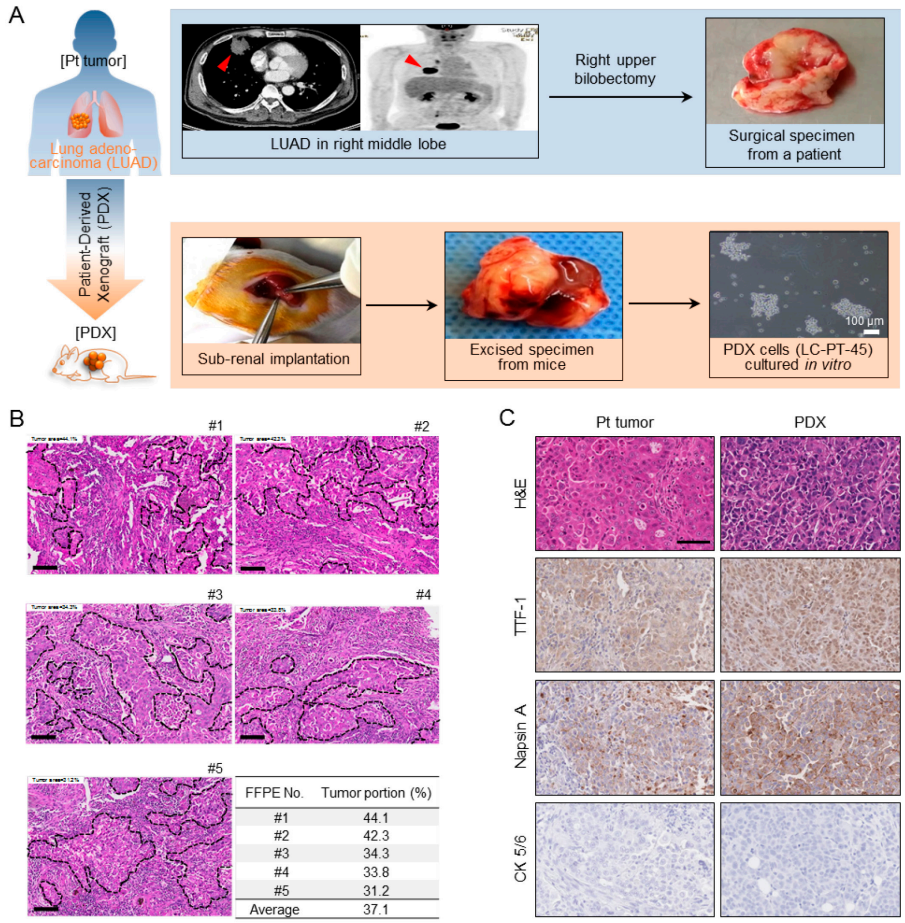


Figure 3. Enriched signatures of cancer cells in PDX. (A) Estimated cancer cell fraction in Pt tumor and PDX cells. The fraction was quantified by histopathological examination (stripe), or estimated based on computational analysis using expression profiles (blue) or WES data (green). (B) Estimated degree of normalized copy number changes in log2 ratio to matched peripheral blood for deletion (green) or amplification (red) are indicated. Representative sites of copy number changes in LUAD are labeled on the right side. (C) Distribution of variant allele frequencies (VAF) of the non-synonymous somatic mutations that overlap between Pt tumor and PDX cells. Color-scaled density map indicates the number of mutations.



FFPE No.	Tumor portion (%)
#1	44.1
#2	42.3
#3	34.3
#4	33.8
#5	31.2
Average	37.1

Figure 4. Propagation of LUAD tumor cells in the xenograft model.

(A) A summarized depiction of the experimental process of tumor engraftment from a LUAD patient into mice. (B) Histological examination by a licensed pathologist determined the tumor area (dotted lines) in FFPE samples of a patient tumor. (C) Evaluation of propagation of LUAD from a patient and in mice by immunohistochemistry analysis, using lung adenocarcinoma cell specific markers (TTF-1 and Napsin A) and a lung squamous cell carcinoma specific marker (CK 5/6). (B-C) Scale bar, 100 μm.

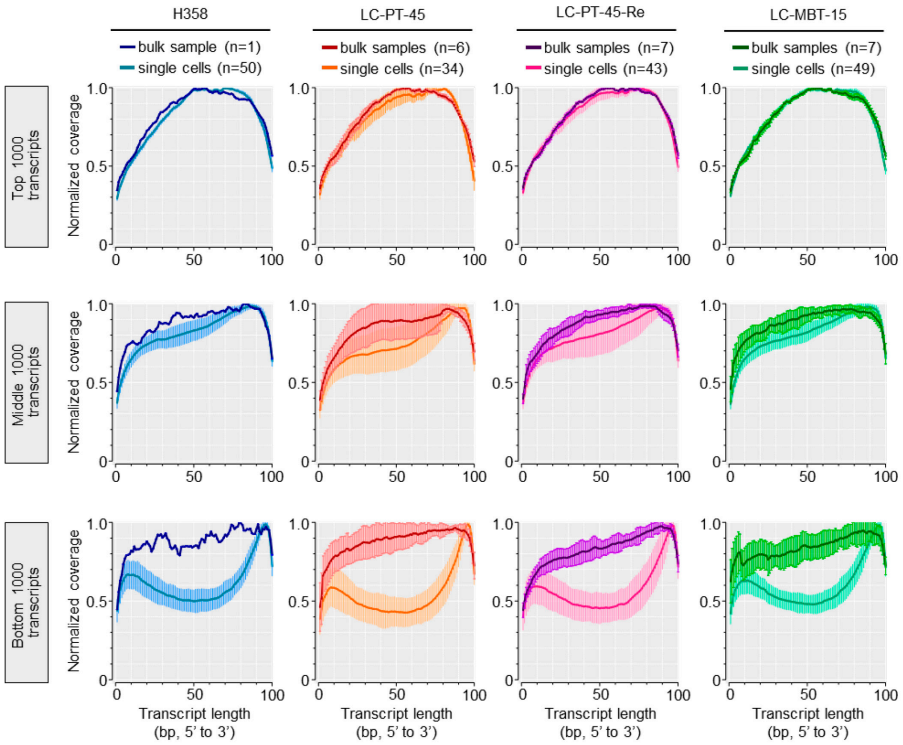


Figure 5. Coverage plots of transcripts based on expression level.

Expression levels of the transcripts were rank-ordered and classified in each sample. Top: 1000 transcripts. Middle: 500 up- and 500 down-transcripts from the median, rank-ordered. Bottom: 1000 transcripts. Coverage ratio was normalized to the maximal degree of coverage in each sample. Standard deviation across samples is depicted as thinner vertical lines over thicker curves.

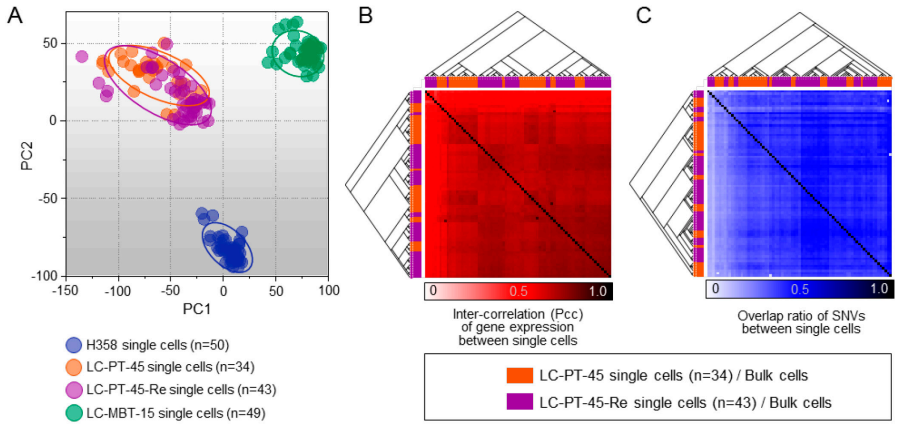


Figure 6. Evaluation of batch effects using a technical replicate set.

(A) Principal component analysis for total data sets of single cells used in this study. (B-C) Interrelation between single cells from LC-PT-45 and LC-PT-45-Re, a technical replicate set, in gene expression (measured by Pcc) (B), and in expressed SNVs (measured by overlap ratio) (C). Unsupervised hierarchical clustering trees were constructed by applying Euclidean distance.

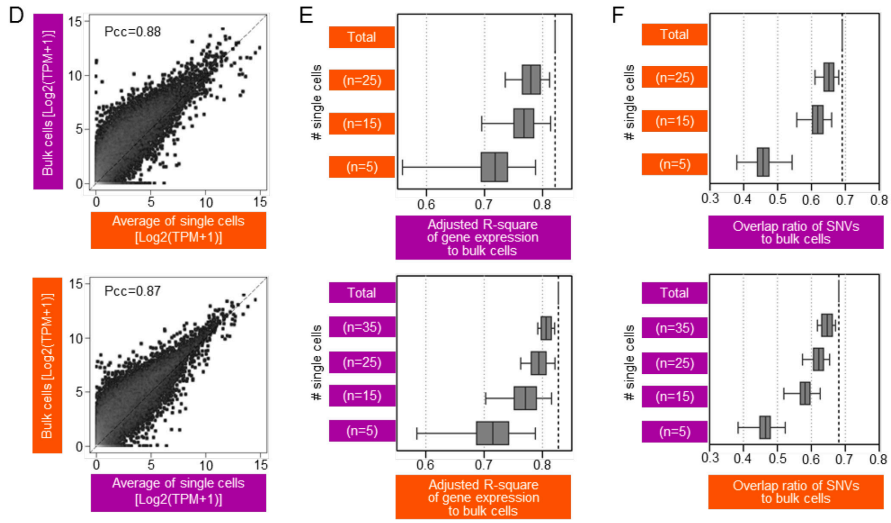


Figure 6. Evaluation of batch effects using a technical replicate set. (D-F) Reciprocal relations between single cells and bulk cells from the other batch set. (D) Scatter plots depicting average gene expression of single cells and bulk cells. The $x=y$ lines (black dotted) with correlation coefficients (Pearson's) for linear fit are shown in each panel. (E) Explanatory power (adjusted R-square) in gene expression of various numbers of single cells towards the bulk cells was determined by multiple regression analysis using randomly selected cell numbers with permutation (×1000). (F) Overlap ratio of expressed SNVs of various single-cell numbers relative to that of the bulk cells was calculated with a randomly selected given number of cells with permutation (×1000). Boxplots in (E) and (F), box=interquartile range (IQR) between the first and the third quartiles, error bars=10^{th}–90^{th} percentiles.

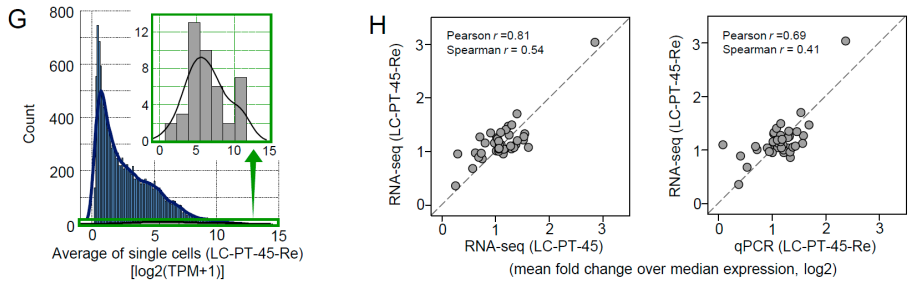


Figure 6. Evaluation of batch effects using a technical replicate set. (G) Distribution of mean expression across single cell RNA-seq data for the total genes (main graph) and for the genes used in qPCR (inset, n=43) (H) Evaluation of gene expression variation across single cells between two batch sets of RNA-seq (left), and between the two technical platforms of RNA-seq and qPCR (right). For parallel comparison (left and right panels), 43 target gene probes were selected for validation. The *x*=*y* lines (black dotted) with correlation coefficients (Pearson *r* and Spearman *r*) for linear fit are shown in each panel.

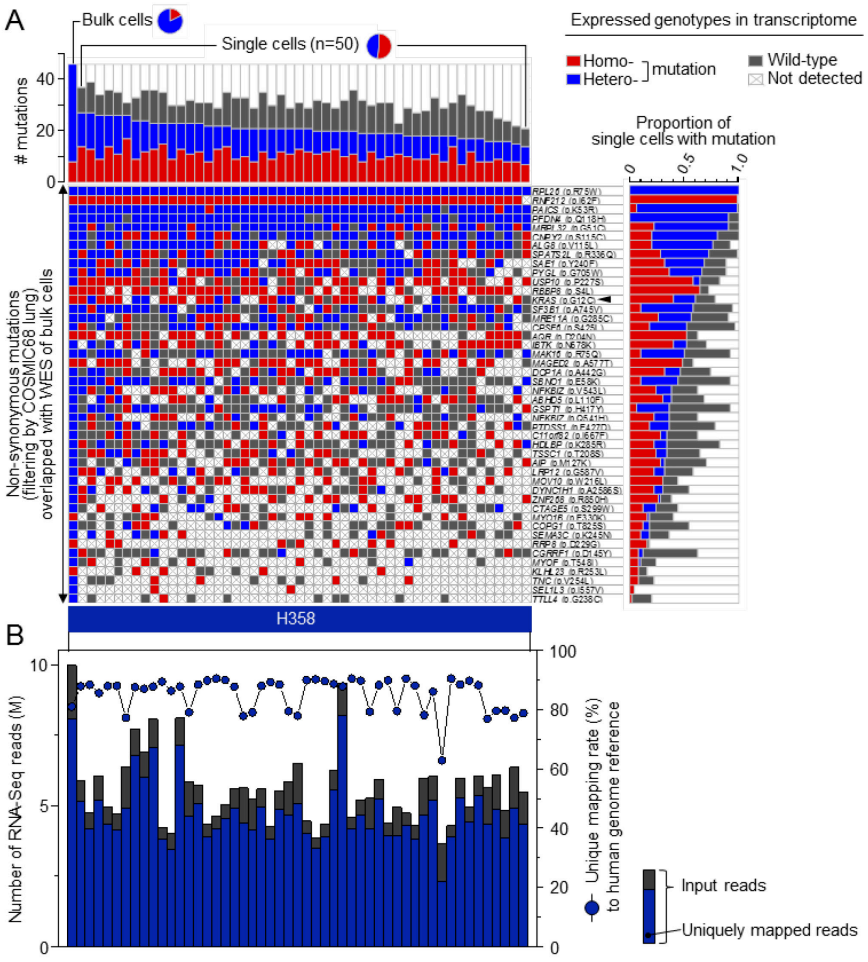


Figure 7. Expressed genotypes of SNVs in H358 cells. (A) Top vertical bars=mutation events per sample; middle heat map=mutation profiles across samples; right horizontal bars=normalized mutation fraction over total single-cells (n=50). (B) Mapping information from RNA-seq reads to a human reference genome (hg19). Vertical bar plots of the number of RNA-seq reads (left y-axis) and scatter plots with a connecting line for the unique mapping rate (uniquely mapped reads/input reads, right y-axis) are in the same order as in (A).

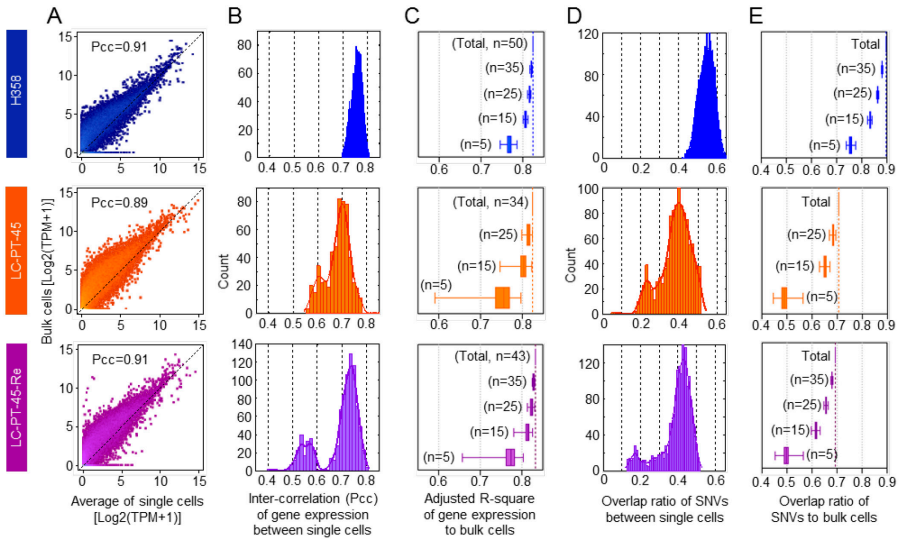


Figure 8. Intratumoral heterogeneity of PDX cells. (A) Scatter plots of the average gene expression of single cells (H358, n=50; LC-PT-45, n=34; LC-PT-45-Re, n=43) compared to those of the corresponding bulk cells (~1×10^5 cells). The $x=y$ lines (black, dotted) with correlation coefficients (Pearson's) for linear fit are shown in each panel. (B) Inter-correlation (Pcc) between gene expressions in single-cells. Density plots were constructed with a kernel function fitting over the histograms. (C) Explanatory power (adjusted R-square) in gene expression of various numbers of single cells towards the bulk cells was determined by multiple regression analysis with randomly selected cell numbers with permutation (×1000). (D) Overlap ratio of expressed SNVs among single-cells. Density plots were constructed with a kernel function fitting over the histograms. (E) Overlap ratio of expressed SNVs of various single-cell numbers relative to that of the bulk cells was calculated with a randomly selected given number of cells with permutation (×1000). Boxplots in (C) and (E), box=interquartile range (IQR) between the first and the third quartiles, error bars=10^{th}–90^{th} percentiles.

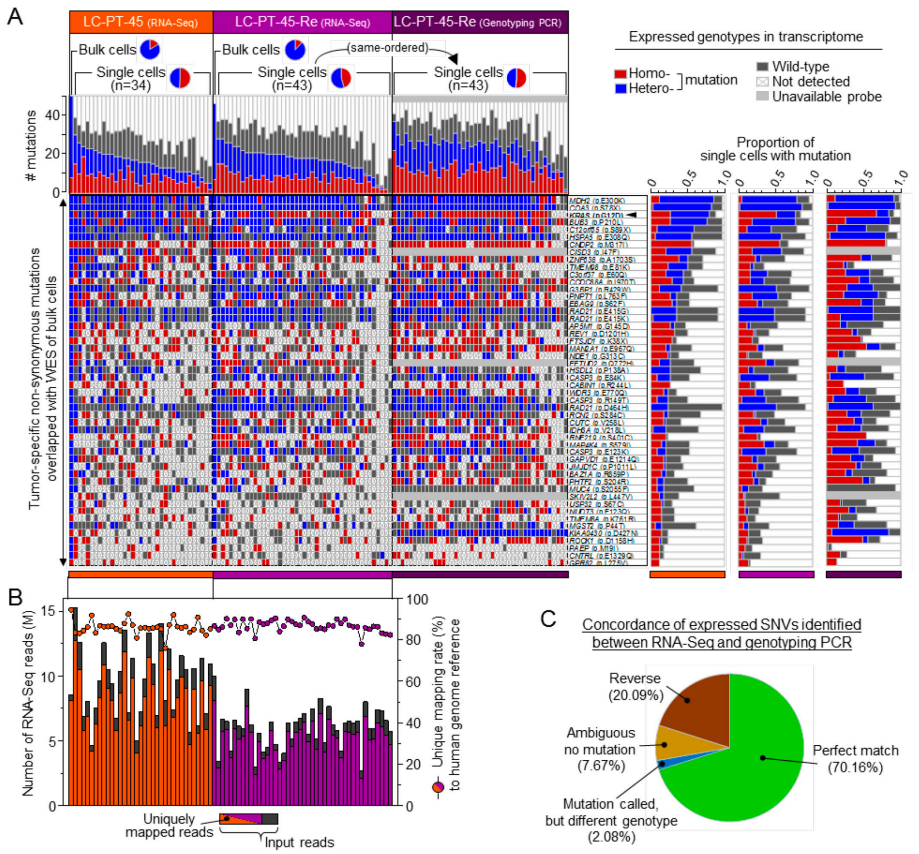


Figure 9. Heterogeneous expression patterns of SNVs in PDX cells. (A) Expressed, tumor-specific, non-synonymous somatic mutations found in more than three single cells of LC-PT-45. The replicate batch (LC-PT-45-Re) of single-cell RNA-seq and that of genotyping PCR are shown together. Top vertical bars=mutation events per sample; middle heat map=mutation profiles across samples; right horizontal bars=normalized mutation fraction over total single-cells (LC-PT-45, n=34; LC-PT-45-Re, n=43). (B) Mapping information from RNA-seq reads to the human reference genome (hg19). Vertical bar plots for the number of RNA-seq reads (left y-axis) and scatter plots with a connecting line for the uniquely mapping rate (uniquely mapped reads/input reads, right y-axis) are in the same order as in (A). (C) Summary of results for the matched samples and the validated targets between RNA-seq and genotyping PCR shown in (A). See Figure 10 for the details.

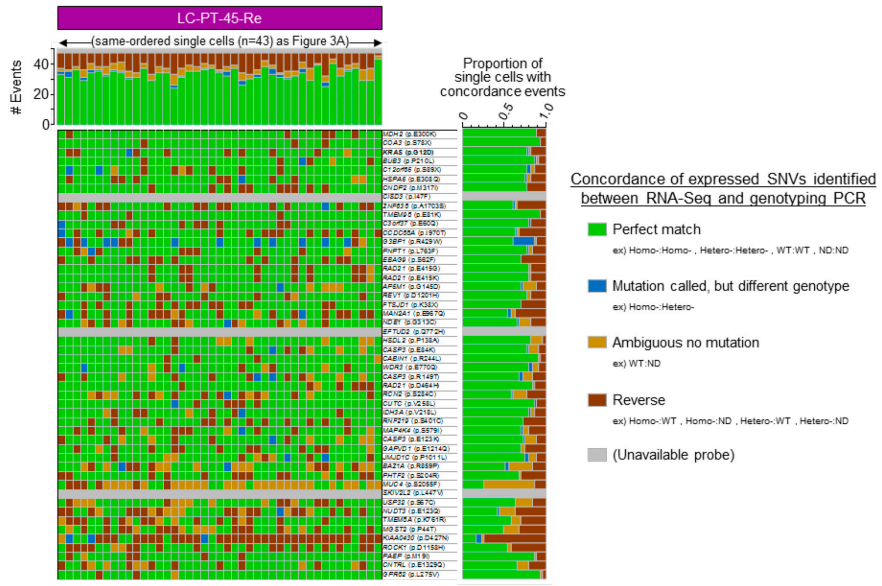


Figure 10. Summary heatmap identifying concordance between RNA-seq and genotyping PCR across matched single cells. Top vertical bars=concordance events per sample; middle heat map=concordance profiles across samples; right horizontal bars=normalized concordance fraction over total single-cells (LC-PT-45-Re, n=43).

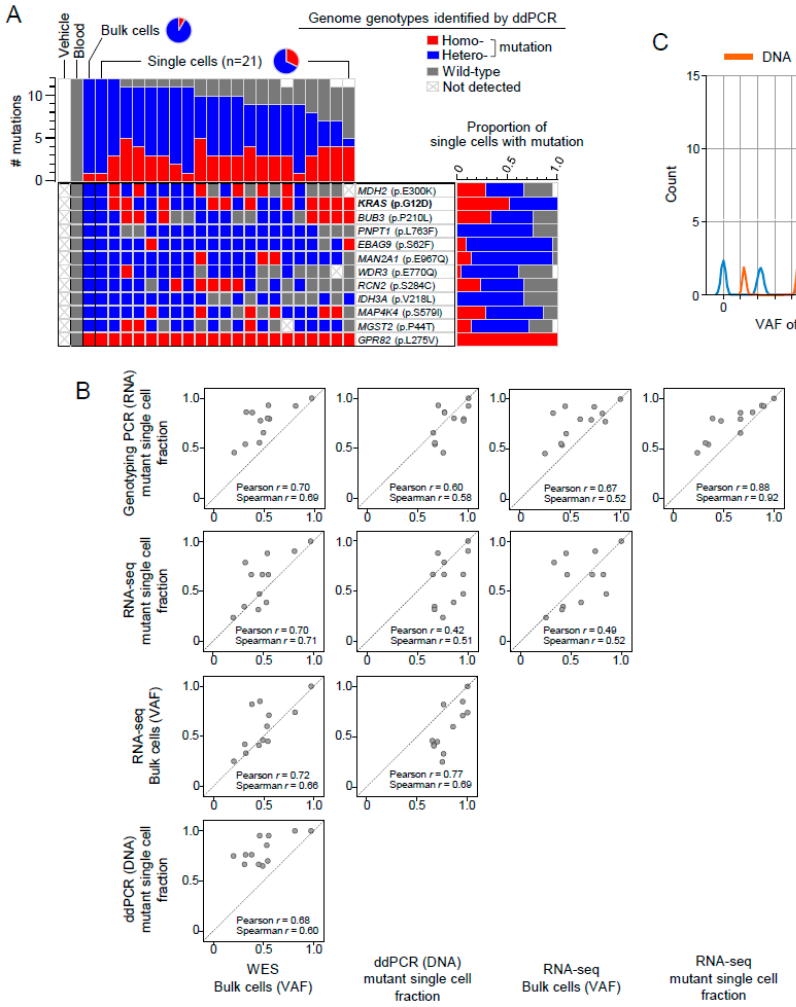


Figure 11. Comparison of various platforms for detecting mutant single cell fractions and variant allele frequencies of bulk cells. (A) The summarized results of ddPCR for selected SNVs at the DNA level. Top left: bar graph of mutation events per sample. Bottom left: heat map of mutation profiles across samples. Right: bar graph of normalized mutation fraction over total single cells (LC-PT-45, n = 21). (B) Multidimensional scatter plots of the comparative fraction of SNVs across various platforms. Black dotted lines are x=y lines with correlation coefficients (Pearson r and Spearman r) for linear fit. (C) The variant allele frequency (VAF) of KRASG12D across single cells separately measured for DNA (by ddPCR) and RNA (by RNA-seq).

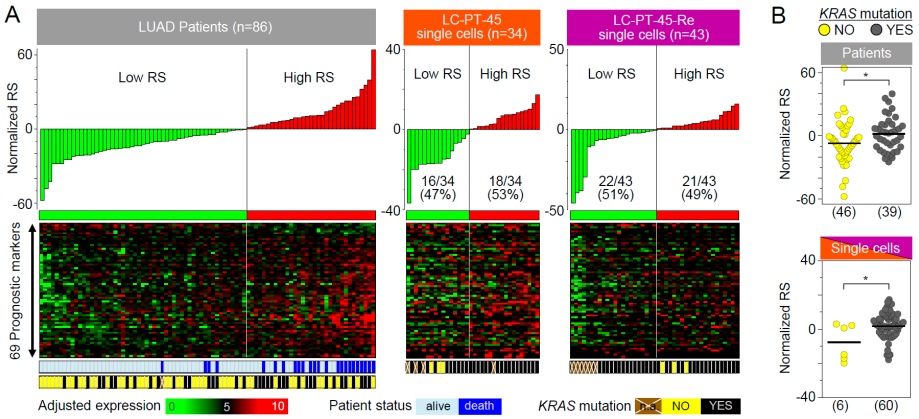


Figure 12. Identification of PDX cell subclones using single-cell RNA-seq data. (A) Top: normalized RS. Middle: heatmaps of expression of 69 prognostic markers. Bottom: KRAS mutation status of each patient (training set, n = 86) or single cell (LC-PT-45, n = 34; LC-PT-45-Re, n = 43). (B) Scatter plots demonstrating the effect of the KRAS mutation on the RSs of LUAD patients and PDX single cells. Horizontal lines represent the mean. *P < 0.05; **P < 0.01.

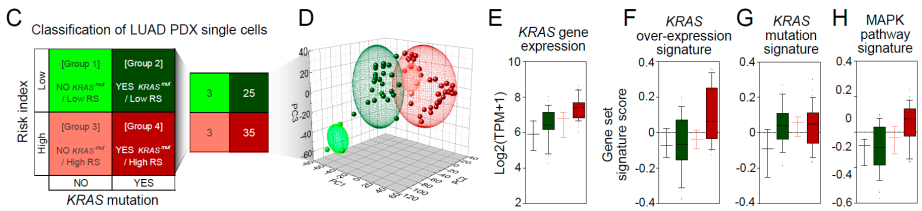


Figure 12. Identification of PDX cell subclones using single-cell RNA-seq data. (C) Semi-supervised clustering of single cells into four groups with estimated RS and KRAS mutant status. (D) Principal component analysis of the genes discriminating the subgroups. Ellipsoids were generated with standard deviations around each group. (E-H) Comparative features among the classified single cell subgroups. (E) KRAS gene expression (Log2 ratio of transcripts per million + 1). Gene set signature scores (computed by gene set variation analysis) corresponding to the KRAS over-expression signature (77) (F), KRAS mutation signature (78) (G), and MAPK pathway signature (gene sets from BioCarta) (H). For the boxplots in (E–H), boxes = the interquartile range between the first and third quartiles, and error bars = 10th–90th percentiles

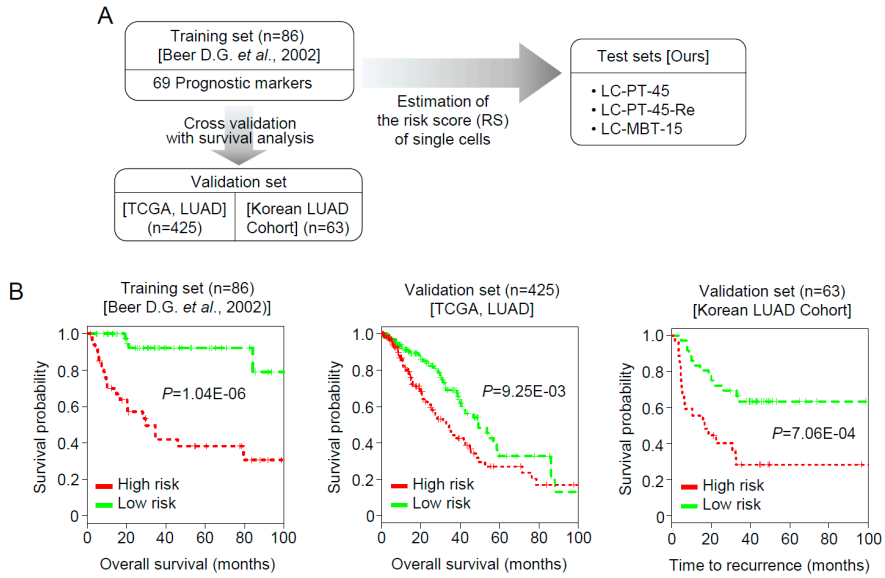


Figure 13. Application of risk scores to patient survival in LUAD cohorts. (A) Strategy to classify single cells according to prognostic marker expression. (B) Kaplan-Meier curves of overall survival of patients in two independent LUAD cohorts and of recurrence-free survival of patients in a Korean LUAD cohort, according to the estimated risk scores (log-rank test).

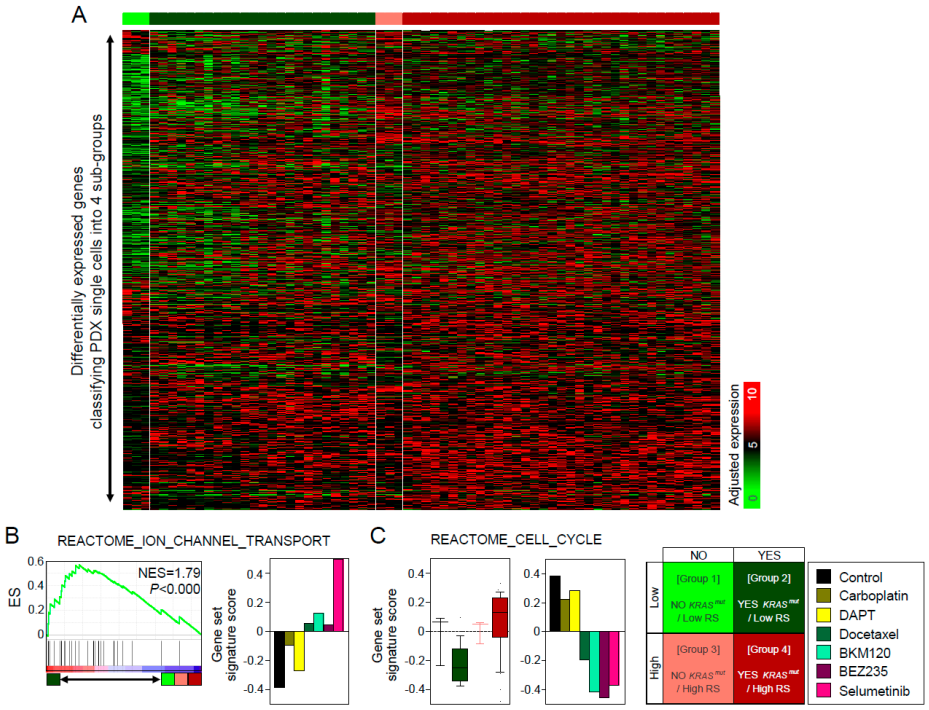


Figure 14. Distinct gene expression signatures among the classified single cell subgroups along with the drug treatment groups. (A) Expression heatmap discriminating single cells into subgroups classified as in Fig. 12C. (B) REACTOME-defined ion channel transport is significantly activated in group 2 compared with the other groups, as determined by gene set enrichment analysis. Statistical significance was determined using the nominal P values. ES enrichment score; NES normalized enrichment score. Gene set activation signatures were estimated for the control and drug-treated PDX cells by gene set variation analysis. (C) Gene expression signature for the cell cycle was estimated by gene set variation analysis. The gene set for the cell cycle signature was obtained from REACTOME.

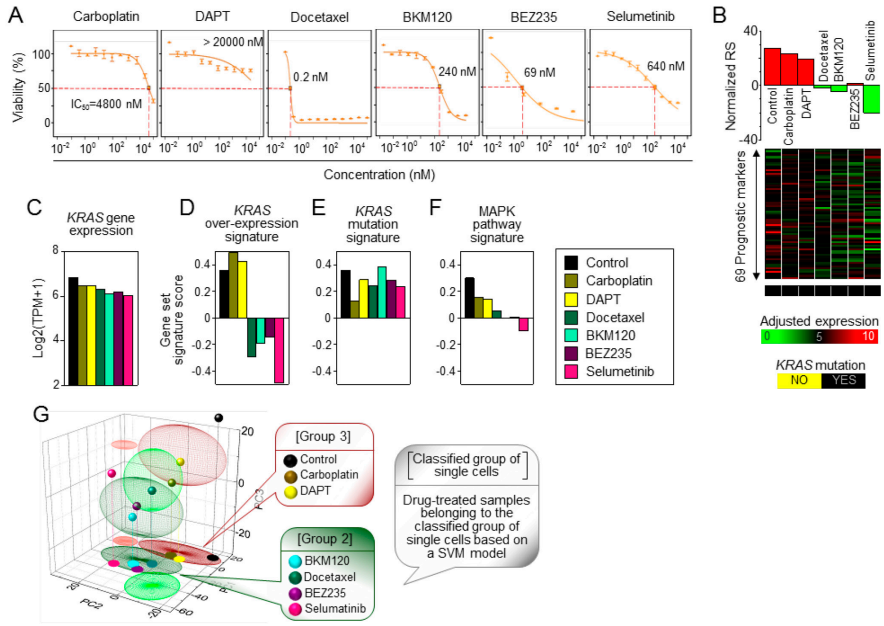
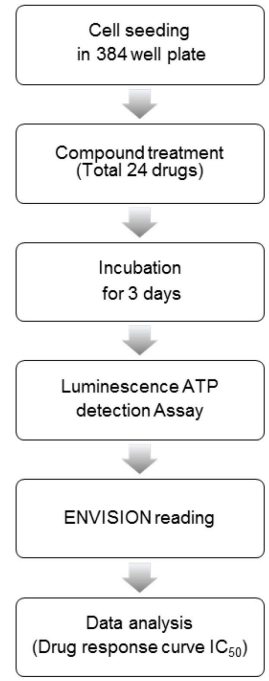


Figure 15. Interpretation of drug responses using single-cell signatures. (A) Dose response curves for the four selected anti-cancer compounds (cytotoxic: Carboplatin, Docetaxel; molecular targeting: DAPT, BKM120, BEZ235, Selumetinib). (B) Normalized RSs (upper) and adjusted-expression of the 69 prognostic markers (middle) with *KRAS* mutant expression (lower) for the control and drug-treated PDX cells. (C-F) Comparative features among the control and drug-treated PDX cells. (C) *KRAS* gene expression (Log2 ratio of TPM+1). Gene set signature scores (computed by Gene Set Variation Analysis) corresponding to the *KRAS* over-expression signature (77) (D), *KRAS* mutation signature (78) (E), and MAPK pathway signature (gene sets from BioCarta) (F). (G) Results from the principal component analysis on single cells along with the control and drug-treated PDX cells. Ellipsoids corresponding to the single cell subgroups [Group 1 (light green), Group 2 (dark green) and Group 3 (dark red)], with the control and drug-treated PDX cells were projected on the PC1-PC2 plane. Using single cell subgroups as a training set, C-SVM classification was applied to a test set of the control and drug-treated PDX cells.

A



B

Drug	Target	IC_{50} (nM)
BEZ235*	*PI3K, mTOR*	69
BKM120*	*PI3K*	240
Cabozatinib	*VEGFR2*	2,900
Carboplatin†	DNA synthesis	4,800
Crizotinib	*C-Met/ALK*	730
DAPT*	γ-secretase	> 20,000
Docetaxel†	microtubule	0.2
Dovitinib	*FLT3/c-Kit, FGFR1/3, VEGFR1-4*	750
Erlotinib	*EGFR*	3,100
Etoposide	topoisomerase II	440
Gefitinib	*EGFR*	1,600
Gemcitabine	DNA synthesis	0.5
Irinotecan	topoisomerase I	230
Lapatinib	*EGFR/HER2*	2,700
Nintedanib	*VEGFR1/2/3, FGFR1/2/3, PDGFRα/β*	> 20,000
Paclitaxel	microtubule	1.5
Pazopanib	*VEGFR1/2/3, PDGFR, FGFR, c-Kit, c-Fms*	6,500
Pemetrexed	folate and metabolite	340
Saracatinib	*SRC, ABL*	2,700
Selumetinib*	*MEK*	640
Sorafenib	multi-targeted RTK (*RAF-1/B-RAF/VEGFR2*)	2,100
Sunitinib	multi-targeted RTK (*VEGFR2/PDGFRβ*)	2,400
Temsirolimus	*mTOR, HIF-1/2, VEGF*	7,800
Vemufafenib	*B-RAF* (p.V600E)	3,000
Vinblastine	microtubule	0.5

Figure 16. Procedure and the results of drug screening for LC-PT-45. (A) Flow charts showing the overall process from PDX cell preparation to drug screening. (B) Summarized list of drugs used in the screening, their known targets, and calculated IC_{50}. The six anti-cancer compounds (†cytotoxic: Carboplatin, Docetaxel; *molecular targeting: DAPT, BKM120, BEZ235, Selumetinib) selected in this study.

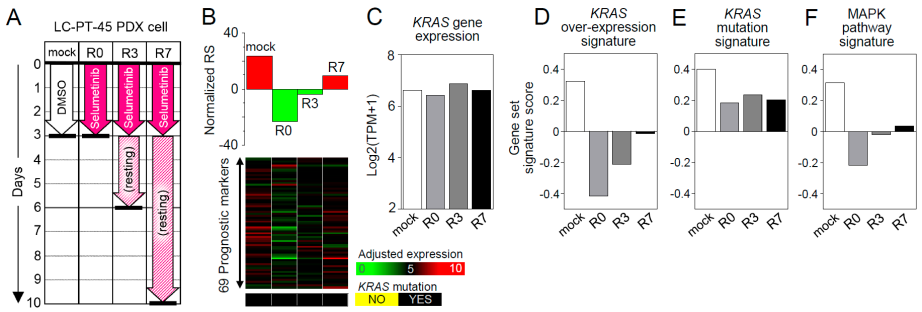


Figure 17. Assessment of phenotypic reversibility for selumetinib-mediated gene expression signatures. (A) The experimental design to examine the change of gene expression under selumetinib. LC-PT-45 PDX cells were serially collected before and after 3-day exposure to 1 μM selumetinib, and on 3 days (R3) and 7 days (R7) after the washout of the drug. (B) Normalized RSs (top) and adjusted-expression of the 69 prognostic markers (middle) with KRAS mutant expression (bottom) for the mock- and selumetinib-treated PDX cells. (C–F) Comparative features among the mock- and selumetinib-treated PDX cells. (C) KRAS gene expression (Log2 ratio of TPM + 1). Gene set signature scores (computed by gene set variation analysis) corresponding to the KRAS overexpression signature (77) (D), KRAS mutation signature (78) (E), and MAPK pathway signature (gene sets from BioCarta) (F).

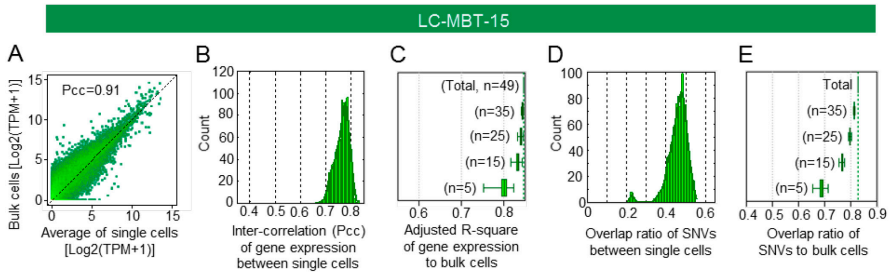


Figure 18. Validation of analytical procedures on an additional PDX, LC-MBT-15. (A) A scatter plot of the average gene expression of single cells (n=49) and that of the corresponding bulk cells (~1×10^5 cells). The $x=y$ lines (black, dotted) with correlation coefficients (Pearson's) for linear fit are shown in each panel. (B) Inter-correlation (Pcc) between gene expressions of single cells. Density plots were constructed with a kernel function fitting over the histograms. (C) Explanatory power (adjusted R-square) in gene expression of various numbers of single cells towards the bulk cells was determined by multiple regression analysis using randomly selected cell numbers with permutation (×1000). (D) Overlap ratio of expressed SNVs among single-cells. Density plots were constructed with a kernel function fitting over the histograms. (E) Overlap ratio of expressed SNVs of various single-cell numbers relative to that of the bulk cells was calculated with a randomly selected given number of cells with permutation (×1000). Boxplot in (E), box=interquartile range (IQR) between the first and the third quartiles, error bars=10^{th}–90^{th} percentiles.

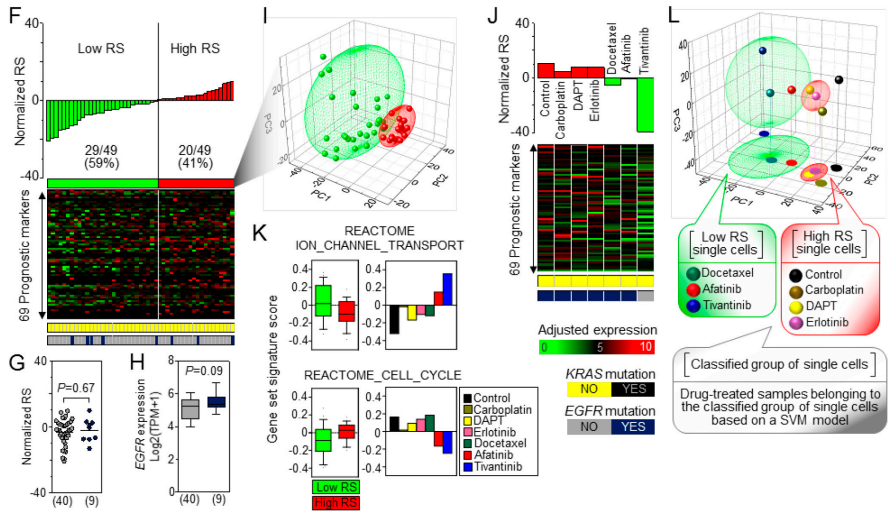


Figure 18. Validation of analytical procedures on an additional PDX, LC-MBT-15. (F) Normalized RS score (upper bar); expression of 69 prognostic markers (middle heatmap); *KRAS* and *EGFR* mutation status (lower bar) of single cells are illustrated. (G) Scatter plots demonstrating the lack impact of the *EGFR* mutation on RSs of LC-MBT-15 single cells. Horizontal lines represents the mean. (H) *EGFR* gene expression (Log2 ratio of TPM+1). Boxplots in (G-H), Box=IQR between the first and the third quartiles, error bars=10th–90th percentiles. (I) Graphical illustration of principal component analysis of the genes discriminating between the low-RS and high-RS subgroups. Ellipsoids were generated with standard deviations around each subgroups. (J) Normalized RSs (upper) and adjusted-expression of the 69 prognostic markers (middle) with *KRAS* and *EGFR* mutation status (lower) for the control and drug-treated PDX cells. (K) Gene set activation signatures were estimated for single cells (left) and the control and drug-treated PDX cells (right) by GSVA. Gene expression signatures for ion channel transport and cell cycle were from REACTOME. (L) Results from the principal component analysis on single cells along with the control and drug-treated PDX cells. Ellipsoids corresponding to the single cell subgroups [Low-RS (green), High-RS (red)], with the control and drug-treated PDX cells are projected on the PC1-PC2 plane. Using single cell subgroups as a training set, C-SVM classification was applied to a test set of the control and drug-treated PDX cells.

Drug	Target	IC_{50} (nM)
Afatinib*	*EGFR/HER2*	270
BKM120	*PI3K*	1000
Cabozantinib	*VEGFR2*	> 20000
Carboplatin†	DNA synthesis	> 10000
Crizotinib	*C-Met/ALK*	1900
DAPT*	γ-secretase	> 20000
Docetaxel†	microtubule	0.098
Dovitinib	*FLT3/c-Kit, FGFR1/3, VEGFR1-4*	1700
Erlotinib*	*EGFR*	> 20000
Etoposide	topoisomerase II	> 10000
Everolimus	*mTOR*	4100
Foretinib	*C-Met/VEGFR-2*	1200
Gefitinib	*EGFR*	7900
Gemcitabine	DNA synthesis	> 10000
Irinotecan	topoisomerase I	16000
Lapatinib	*EGFR/HER2*	3400
Nintedanib	*VEGFR1/2/3, FGFR1/2/3, PDGFRα/β*	2100
Paclitaxel	microtubule	2.1
Pazopanib	*VEGFR1/2/3, PDGFR, FGFR, c-Kit, c-Fms*	14000
Pemetrexed	folate and metabolite	610
Selumetinib	*MEK*	1500
Sorafenib	multi-targeted RTK (*RAF-1/B-RAF/VEGFR2*)	4000
Sunitinib	multi-targeted RTK (*VEGFR2/PDGFRβ*)	3900
Temsirolimus	*mTOR, HIF-1/2, VEGF*	570
Tivantinib*	*MET*	370
Vandetanib	*EGFR/VEGF/RET*	5700
Vemurafenib	*B-RAF* (p.V600E)	9200
Vinblastine	microtubule	0.8

Figure 19. The results of drug screening for LC-MBT-15. Summarized list of drugs used in the screening, their known targets, and calculated IC_{50}. The six anti-cancer compounds (†cytotoxic: Carboplatin, Docetaxel; *molecular targeting: Afatinib, DAPT, Erlotinib, Tivantinib) selected in this study.

Table 1. Somatic mutations identified both in patient tumor and PDX pooled cells.

Gene	RefSeq	Chr	Position	Genome change	cDNA change	Protein change	Variant	Pt_tumor (VAF)	Xeno_pooled (VAF)
PRAMEF6	NM_001010889	1	13002278	g.chr1:13002278G>A	c.C71T	p.A24V	nonsynonymous	0.09	0.21
PRAMEF14	NM_001099854	1	13671690	g.chr1:13671690A>C	c.T2G	p.M1R	nonsynonymous	0.22	0.67
CROCC	NM_014675	1	17266536	g.chr1:17266536G>C	c.G1756C	p.D586H	nonsynonymous	0.43	0.72
DEPDC1	NM_001114120	1	68948323	g.chr1:68948323C>G	c.G1168C	p.V390L	nonsynonymous	0.53	0.52
SRGAP2D	NM_001271887	1	121116733	g.chr1:121116733A>G	c.A290G	p.Q97R	nonsynonymous	0.12	0.36
NOTCH2NL	NM_203458	1	145281633	g.chr1:145281633C>A	c.C563A	p.P188H	nonsynonymous	0.21	0.09
NOTCH2NL	NM_203458	1	145281656	g.chr1:145281656A>T	c.A586T	p.T196S	nonsynonymous	0.21	0.08
NBPF10	NM_001039703	1	145293515	g.chr1:145293515A>G	c.A110G	p.N37S	nonsynonymous	0.26	0.18
NBPF10	NM_001039703	1	145299808	g.chr1:145299808T>C	c.T857C	p.M286T	nonsynonymous	0.43	0.13
NBPF10	NM_001039703	1	145323656	g.chr1:145323656A>T	c.A3718T	p.I1240F	nonsynonymous	0.25	0.14
IRF2BP2	NM_001077397	1	234745009	g.chr1:234745009A>G	c.T232C	p.S78P	nonsynonymous	0.63	0.58
OR2T34	NM_001001821	1	248737511	g.chr1:248737511A>G	c.T548C	p.F183S	nonsynonymous	0.60	0.24
OR2T34	NM_001001821	1	248737664	g.chr1:248737664C>T	c.G395A	p.C132Y	nonsynonymous	0.30	0.22
MST1L	NM_001271733	1	17084304	g.chr1:17084304C>T	c.G1713A	p.W571X	stopgain	0.11	0.10
EHBP1	NM_015252	2	63264728	g.chr2:63264728G>C	c.G3375C	p.K1125N	nonsynonymous	0.03	0.35
CSRNP3	NM_001172173	2	166535524	g.chr2:166535524G>C	c.G1019C	p.G340A	nonsynonymous	0.03	0.33
ZNF804A	NM_194250	2	185803718	g.chr2:185803718C>A	c.C3595A	p.H1199N	nonsynonymous	0.02	0.67
CPS1	NM_001122633	2	211540548	g.chr2:211540548C>A	c.C4276A	p.L1426I	nonsynonymous	0.04	0.68
FOXD4L1	NM_012184	2	114257266	g.chr2:114257266C>T	c.C433T	p.R145C	nonsynonymous	0.29	0.33
POTEJ	NM_001277083	2	131390121	g.chr2:131390121G>C	c.G1190C	p.S397T	nonsynonymous	0.18	0.11
UGT1A8	NM_019076	2	234526871	g.chr2:234526871C>G	c.C518G	p.A173G	nonsynonymous	0.60	0.67
FRG2C	NM_001124759	3	75714041	g.chr3:75714041G>A	c.G250A	p.G84R	nonsynonymous	0.11	0.18
ALDH1L1	NM_012190	3	125874332	g.chr3:125874332C>G	c.G543C	p.R181S	nonsynonymous	0.04	0.46
SLITRK3	NM_014926	3	164907950	g.chr3:164907950G>T	c.C669A	p.S223R	nonsynonymous	0.05	0.51
MUC4	NM_018406	3	195511780	g.chr3:195511780G>A	c.C6671T	p.P2224L	nonsynonymous	0.23	0.22
MST1	NM_020998	3	49726070	g.chr3:49726070G>A	c.C55T	p.P19S	nonsynonymous	0.22	0.36
ZNF717	NM_001128223	3	75786469	g.chr3:75786469G>A	c.C2305T	p.H769Y	nonsynonymous	0.09	0.09
ZNF717	NM_001128223	3	75787918	g.chr3:75787918C>T	c.G856A	p.V286I	nonsynonymous	0.10	0.13
ZNF717	NM_001128223	3	75787926	g.chr3:75787926T>C	c.A848G	p.Y283C	nonsynonymous	0.11	0.13
ZNF717	NM_001128223	3	75790427	g.chr3:75790427C>T	c.G277A	p.A93T	nonsynonymous	0.12	0.14
MUC4	NM_018406	3	195511273	g.chr3:195511273G>A	c.C7178T	p.A2393V	nonsynonymous	0.25	0.20
MUC4	NM_018406	3	195511390	g.chr3:195511390T>G	c.A7061C	p.H2354P	nonsynonymous	0.33	0.38
MUC4	NM_018406	3	195513433	g.chr3:195513433G>A	c.C5018T	p.A1673V	nonsynonymous	0.14	0.19
MUC4	NM_018406	3	195513563	g.chr3:195513563T>C	c.A4888G	p.T1630A	nonsynonymous	0.20	0.09
MUC4	NM_018406	3	195513566	g.chr3:195513566C>G	c.G4885C	p.D1629H	nonsynonymous	0.22	0.13
ZNF518B	NM_053042	4	10447162	g.chr4:10447162G>C	c.C791G	p.S264C	nonsynonymous	0.03	0.45
ZNF518B	NM_053042	4	10447837	g.chr4:10447837T>G	c.A116C	p.E39A	nonsynonymous	0.03	0.47
TIGD2	NM_145715	4	90034567	g.chr4:90034567G>A	c.G442A	p.D148N	nonsynonymous	0.03	0.49
POU4F2	NM_004575	4	147561416	g.chr4:147561416A>T	c.A686T	p.H229L	nonsynonymous	0.04	0.26
USP17L11	NM_001256854	4	9261083	g.chr4:9261083C>T	c.C1234T	p.H412Y	nonsynonymous	0.29	0.15
MROH2B	NM_173489	5	41004889	g.chr5:41004889C>A	c.G3998T	p.G1333V	nonsynonymous	0.04	0.54
PPARGC1B	NM_133263	5	149200045	g.chr5:149200045A>C	c.A128C	p.D43A	nonsynonymous	0.04	0.55
PCDHA4	NM_018907	5	140186980	g.chr5:140186980G>A	c.G208A	p.G70S	nonsynonymous	0.20	0.19
PCDHA4	NM_018907	5	140186984	g.chr5:140186984G>A	c.G212A	p.R71H	nonsynonymous	0.21	0.18
PCDHA4	NM_018907	5	140186986	g.chr5:140186986G>C	c.G214C	p.G72R	nonsynonymous	0.22	0.18
PCDHA4	NM_018907	5	140186990	g.chr5:140186990G>A	c.G218A	p.G73D	nonsynonymous	0.26	0.19
PCDHB10	NM_018930	5	140573754	g.chr5:140573754A>C	c.A1629C	p.R543S	nonsynonymous	0.25	0.24
NOP16	NM_001256539	5	175811095	g.chr5:175811095G>A	c.C586T	p.R196C	nonsynonymous	0.09	0.09
P4HA2	NM_001142599	5	131552905	g.chr5:131552905G>A	c.C316T	p.Q106X	stopgain	0.04	0.39
GRIK2	NM_175768	6	102247637	g.chr6:102247637G>C	c.G1066C	p.G356R	nonsynonymous	0.04	1.00
TNXB	NM_019105	6	32011235	g.chr6:32011235C>T	c.G11623A	p.V3875I	nonsynonymous	0.67	0.47
HSP90AB1	NM_001271969	6	44221316	g.chr6:44221316G>A	c.G2156A	p.R719H	nonsynonymous	0.11	0.15
CTAGE9	NM_001145659	6	132029865	g.chr6:132029865G>A	c.C2293T	p.L765F	nonsynonymous	0.75	0.28
HDAC9	NM_001204144	7	18535887	g.chr7:18535887G>T	c.G88T	p.G30W	nonsynonymous	0.04	0.11
FIGNL1	NM_022116	7	50513734	g.chr7:50513734C>A	c.G1252T	p.V418L	nonsynonymous	0.06	0.20
STAG3	NM_012447	7	99798087	g.chr7:99798087T>A	c.T1782A	p.D594E	nonsynonymous	0.04	0.45
NACAD	NM_001146334	7	45123210	g.chr7:45123210G>A	c.C2569T	p.P857S	nonsynonymous	0.33	0.67
ZNF479	NM_033273	7	57188016	g.chr7:57188016A>G	c.T1106C	p.M369T	nonsynonymous	0.12	0.71
ZNF727	NM_001159522	7	63538334	g.chr7:63538334A>G	c.A907G	p.K303E	nonsynonymous	0.44	0.61
MUC12	NM_001164462	7	100644993	g.chr7:100644993C>T	c.C11149T	p.R3717C	nonsynonymous	0.10	0.10
OR9A2	NM_001001658	7	142724073	g.chr7:142724073A>T	c.T147A	p.C49X	stopgain	0.04	0.61
ARHGEF5	NM_005435	7	144074230	g.chr7:144074230G>A	c.G4478A	p.W1493X	stopgain	0.10	0.11
PSKH2	NM_033126	8	87076831	g.chr8:87076831C>G	c.G215C	p.G72A	nonsynonymous	0.06	1.00
NIPAL2	NM_024759	8	99264759	g.chr8:99264759C>A	c.G308T	p.G103V	nonsynonymous	0.03	0.38
EBAG9	NM_198120	8	110566980	g.chr8:110566980C>T	c.C185T	p.S62F	nonsynonymous	0.05	0.55
BAI1	NM_001702	8	143546093	g.chr8:143546093C>A	c.C534A	p.N178K	nonsynonymous	0.05	0.69
CYP11B2	NM_000498	8	143998620	g.chr8:143998620C>T	c.G250A	p.G84R	nonsynonymous	0.03	0.38
KIFC2	NM_145754	8	145692366	g.chr8:145692366C>G	c.C203G	p.S68W	nonsynonymous	0.06	0.57
WNK2	NM_006648	9	96060259	g.chr9:96060259C>A	c.C5833A	p.R1945S	nonsynonymous	0.06	0.53
OR13D1	NM_001004484	9	107457517	g.chr9:107457517C>G	c.C815G	p.T272S	nonsynonymous	0.07	0.48
LPAR1	NM_057159	9	113704388	g.chr9:113704388G>A	c.C106T	p.R36X	stopgain	0.04	0.45
WDFY4	NM_020945	10	49968444	g.chr10:49968444G>T	c.G2512T	p.V838L	nonsynonymous	0.03	0.67
OTOG	NM_001277269	11	17579843	g.chr11:17579843A>T	c.A1013T	p.Q338L	nonsynonymous	0.03	0.52

OR4C6	NM_001004704	11	55433523	g.chr11:55433523G>T	c.G881T	p.S294I	nonsynonymous	0.03	0.46
KCNK4	NM_033310	11	64067157	g.chr11:64067157C>T	c.C1141T	p.R381C	nonsynonymous	0.04	0.50
FAT3	NM_001008781	11	92531002	g.chr11:92531002G>A	c.G4823A	p.G1608E	nonsynonymous	0.04	0.53
TRPC6	NM_004621	11	101374888	g.chr11:101374888C>A	c.G812T	p.R271I	nonsynonymous	0.03	0.68
SORL1	NM_003105	11	121492874	g.chr11:121492874C>G	c.C6068G	p.S2023X	stopgain	0.04	0.54
KRAS	NM_033360	12	25398284	g.chr12:25398284C>T	c.G35A	p.G12D	nonsynonymous	0.08	0.81
FMNL3	NM_198900	12	50059631	g.chr12:50059631G>A	c.C352T	p.R118C	nonsynonymous	0.06	0.26
KRT81	NM_002281	12	52684024	g.chr12:52684024T>G	c.A416C	p.Q139P	nonsynonymous	0.23	0.29
KRT6C	NM_173086	12	52867233	g.chr12:52867233C>T	c.G289A	p.G97R	nonsynonymous	0.40	0.14
KRT6C	NM_173086	12	52867260	g.chr12:52867260C>T	c.G262A	p.G88R	nonsynonymous	0.33	0.27
PABPC3	NM_030979	13	25670953	g.chr13:25670953G>A	c.G617A	p.R206H	nonsynonymous	0.11	0.18
PABPC3	NM_030979	13	25670955	g.chr13:25670955C>T	c.C619T	p.L207F	nonsynonymous	0.11	0.18
PABPC3	NM_030979	13	25671015	g.chr13:25671015A>G	c.A679G	p.S227G	nonsynonymous	0.12	0.14
PABPC3	NM_030979	13	25671027	g.chr13:25671027A>G	c.A691G	p.K231E	nonsynonymous	0.21	0.22
PABPC3	NM_030979	13	25671089	g.chr13:25671089G>T	c.G753T	p.M251I	nonsynonymous	0.16	0.28
AP5M1	NM_018229	14	57741321	g.chr14:57741321G>A	c.G434A	p.G145D	nonsynonymous	0.03	0.29
CCDC88C	NM_001080414	14	91883108	g.chr14:91883108G>C	c.C135G	p.I45M	nonsynonymous	0.05	0.67
PPP1R13B	NM_015316	14	104212814	g.chr14:104212814T>C	c.A1046G	p.Y349C	nonsynonymous	0.03	0.47
AHNAK2	NM_138420	14	105416323	g.chr14:105416323T>C	c.A5465G	p.K1822R	nonsynonymous	0.30	0.21
CSPG4	NM_001897	15	75981550	g.chr15:75981550C>T	c.G1856A	p.R619Q	nonsynonymous	0.43	0.22
ADAMTS7	NM_014272	15	79083120	g.chr15:79083120G>A	c.C920T	p.T307M	nonsynonymous	0.75	0.70
KIAA0430	NM_014647	16	15725310	g.chr16:15725310C>T	c.G1279A	p.D427N	nonsynonymous	0.09	0.64
FTSJD1	NM_018348	16	71319712	g.chr16:71319712T>A	c.A112T	p.K38X	stopgain	0.04	1.00
TRPV3	NM_145068	17	3421994	g.chr17:3421994A>T	c.T1961A	p.F654Y	nonsynonymous	0.05	0.99
SLC47A2	NM_152908	17	19617253	g.chr17:19617253C>A	c.G328T	p.V110L	nonsynonymous	0.04	0.65
MPO	NM_000250	17	56355311	g.chr17:56355311C>A	c.G1081T	p.G361W	nonsynonymous	0.04	0.47
USP32	NM_032582	17	58379052	g.chr17:58379052G>C	c.C200G	p.S67C	nonsynonymous	0.04	0.42
TBX2	NM_005994	17	59479207	g.chr17:59479207G>T	c.G558T	p.M186I	nonsynonymous	0.04	0.52
FBXW10	NM_001267585	17	18653070	g.chr17:18653070G>A	c.G706A	p.E236K	nonsynonymous	0.25	0.39
TBC1D3	NM_001123391	17	36293042	g.chr17:36293042C>A	c.C1073A	p.P358Q	nonsynonymous	0.08	0.15
CNTNAP1	NM_003632	17	40843834	g.chr17:40843834G>A	c.G2355A	p.W785X	stopgain	0.03	0.49
TMEM259	NM_001033026	19	1011133	g.chr19:1011133T>G	c.A1279C	p.S427R	nonsynonymous	0.05	0.39
CYP4F12	NM_023944	19	15807267	g.chr19:15807267G>C	c.G1342C	p.E448Q	nonsynonymous	0.03	0.62
UNC13A	NM_001080421	19	17749956	g.chr19:17749956G>A	c.C3017T	p.S1006F	nonsynonymous	0.04	0.32
IRGC	NM_019612	19	44223683	g.chr19:44223683G>A	c.G973A	p.D325N	nonsynonymous	0.04	0.52
SHANK1	NM_016148	19	51189562	g.chr19:51189562C>T	c.G2509A	p.E837K	nonsynonymous	0.03	0.48
ZNF208	NM_007153	19	22155725	g.chr19:22155725C>G	c.G2111C	p.W704S	nonsynonymous	0.09	0.12
FCGBP	NM_003890	19	40392585	g.chr19:40392585T>G	c.A7919C	p.E2640A	nonsynonymous	0.45	0.42
FCGBP	NM_003890	19	40392747	g.chr19:40392747A>T	c.T7757A	p.L2586Q	nonsynonymous	0.25	0.14
RAB22A	NM_020673	20	56918784	g.chr20:56918784A>T	c.A127T	p.M43L	nonsynonymous	0.04	0.43
KRTAP10-6	NM_198688	21	46012160	g.chr21:46012160G>C	c.C206G	p.P69R	nonsynonymous	0.63	0.44
KRTAP10-6	NM_198688	21	46012182	g.chr21:46012182G>A	c.C184T	p.R62C	nonsynonymous	0.78	0.71
KRTAP10-12	NM_198699	21	46117285	g.chr21:46117285C>T	c.C169T	p.R57C	nonsynonymous	0.15	0.10
KRTAP10-12	NM_198699	21	46117286	g.chr21:46117286G>A	c.G170A	p.R57H	nonsynonymous	0.15	0.10
CABIN1	NM_012295	22	24447361	g.chr22:24447361G>T	c.G731T	p.R244L	nonsynonymous	0.05	1.00
GPR82	NM_080817	X	41587102	g.chrX:41587102C>G	c.C823G	p.L275V	nonsynonymous	0.05	0.97
PHKA1	NM_002637	X	71870302	g.chrX:71870302C>A	c.G1262T	p.G421V	nonsynonymous	0.08	1.00
NXF5	NM_032946	X	101092779	g.chrX:101092779C>A	c.G894T	p.R298S	nonsynonymous	0.05	1.00
LOC100129520	NM_001195272	X	124455458	g.chrX:124455458T>C	c.T1490C	p.L497P	nonsynonymous	0.04	1.00
LOC100129520	NM_001195272	X	124456210	g.chrX:124456210G>C	c.G2242C	p.G748R	nonsynonymous	0.07	0.97
FAM104B	NM_001166700	X	55172572	g.chrX:55172572G>A	c.C296T	p.T99I	nonsynonymous	0.09	0.11
ARMCX4	NM_001256155	X	100749041	g.chrX:100749041A>G	c.A5465G	p.E1822G	nonsynonymous	0.44	0.67
RBM10	NM_005676	X	47041409	g.chrX:47041409C>T	c.C1753T	p.Q585X	stopgain	0.04	0.99

Table 2. Prognostic genes used for computing risk scores.

Gene	coefficient-beta	Gene	coefficient-beta
ADM	0.495	KRT7	0.542
AGFG1	0.517	KYNU	0.335
AKAP12	0.223	LAMB1	0.344
ALDOA	0.730	MT2A	0.001
ATP2B1	0.265	NACA	1.090
BAG1	0.340	NME2	0.528
BBS9	-0.444	P2RX5	-0.635
CASP4	0.472	PDAP1	0.702
CDS1	0.202	PEX7	1.158
CKAP4	0.687	PNP	0.406
CRK	0.825	POLD3	0.514
CSTB	0.434	PPIF	0.419
CYP24A1	0.229	PRDM2	-0.362
DBP	-0.607	RELA	0.400
DEFB1	0.204	RND3	0.262
EIF1	1.342	RPS26	0.435
ERBB2	0.507	RPS3	1.054
FADD	0.932	S100P	0.194
FEZ2	0.494	SERPINE1	0.242
FUCA1	-0.523	SLC20A1	0.258
FURIN	0.294	SLC2A1	0.513
FUT3	0.284	STARD3	0.127
GAPDH	1.199	STC1	0.318
GARS	0.318	STX1A	0.405
GRB7	0.388	TMF1	1.067
H2AFZ	0.579	TPBG	0.708
HLA-B	-0.328	TRA2A	-0.773
HMBS	0.524	TUBA1A	-0.130
HPCAL1	0.318	UBC	-0.025
ITGA2	0.222	UGP2	0.935
KIAA0020	0.507	UQCRC2	0.591
KLF10	0.606	VDAC2	0.426
KLF6	0.421	VEGFA	0.548
KRT18	0.773	WNT10B	0.606

Table 3. Information on primers used in qPCR (expression).

gene	target transcript coordinates	Forward primer	Reverse primer
AGFG1	chr2:227498000-227501214	AAAAGCCACGCCCACCATTA	ACAGCAGCCTCAGACTCCA
AKAP12	chr6:151350000-151354500	GGTTATAAGGCTTGCACCTTCA	AGGTGTTCTGTGGTCTGGAA
ALDOA	chr16:30065500-30070000	GAAAGGCTGAGGCAGGAGAATA	AGTGCAGTGGCATGATCTCA
ATP2B1	chr12:90000000-90004500	GGCTGTTCTTGAACACCTGAC	ATGATGGCTCACGCCTGTAA
BAG1	chr9:33255525-33260441	TTCCCTGGGTCTGTATCCTGTA	TGGCCCAGACATTGAGAACA
CASP4	chr11:104945500-104950000	AGCCGGCTGTGAGTCAATTA	TTCACCTTGCCCACTGCTTA
CKAP4	chr12:106240000-106244500	AGTTGTTCATCCGTGCTTCC	AGGTGCTTCAAGAACGTGAC
CRK	chr17:1423000-1427500	TTGGCCAGGCTGATCTTGAA	TCACGCCTGTAATCCAAGCA
CSTB	chr21:43772512-43776445	AAAGCAGCTGGAGGAAAAGAC	TACGGCCACATTGGGGATTA
EIF1	chr17:41688935-41690999	CATGGACTCTGCACCTTTTCAC	TTGCGGATCAAGAAGCTCCA
ERBB2	chr17:39700000-39704500	ACAGGGAAAGCTGTGGGAAA	TACGCCTCCAACACACTGAA
GAPDH	chr12:6534534-6538359	CACAGTGGCTCATGCTTGTA	GCTGGTCTCAAACTCCTGAC
GARS	chr7:30610000-30614500	GGGATGTTCAGGACAAACCA	ACACCAAGAACCTGGATGAC
H2AFZ	chr4:99948088-99950388	TTAGGCCTGCAGAACAGACA	ACTGAACTCATACCGCAGGAA
HLA-B	chr6:31353872-31357187	TCAAGTTCTCTTCCTCCCCAAC	TGCTCCAGCATCTACAGCAA
HPCAL1	chr2:10424347-10427352	TCCTGATCAAAGCACCTCCA	TTGGTGTTCTGGGGAACTCA
ITGA2	chr5:53000000-53004500	ACCAGTCCCAGTGAGATGAA	CCAGCATGATCAGTGCAGAA
KIAA0020	chr9:2810000-2814500	AGTTCAGGGATGCCAGGAA	GGGAGAGAAGGTAGGCTATGAA
KLF10	chr8:102649000-102653000	GCTACAGCAGCAGAGTTGACTA	CTGTAAAGGGGCAGGCAGTAA
KLF6	chr10:3778000-3782500	CACACACACACACACACACA	TGTTCTCTTGGGACGACTTGAA
KRT18	chr12:52948871-52952900	TCCCATGTCCCAGTCAATTCC	TACCTGGGAGGGGATGTTCA
KRT19	chr17:41523617-41528308	GGAGGTGTCATTGGAGCTGAA	AGCAGCTTCCACCACTTCAA

KRT7	chr12:52238500-52243000	TATGCAGACTGCCTGGCTAA	ACCTTGTGGGTGCTCCATAA
KYNU	chr2:143000000-143004500	TTTGGGTGGCTTCTCTTCCA	ACTAACAGCTGGTATGCCTGAA
MT2A	chr16:56608199-56609497	CCATTGCTTCTTGGATTCCC	ATTGCTTGAGGATGTACTCCA
NACA	chr12:56714000-56718600	TGTGGCTCACGCCTGTAAA	TTGGCCAGGTTGGTCTTGAA
NME2	chr17:51166778-51171747	AGTGCCCACTCTGTGTTTCA	GGGAGAAAACTGCAACGTGAA
PDAP1	chr7:99397500-99402000	CCAGTTGTTTGCTTCGGCTA	GGACCTGATGGAGCTGAAAAC
PNP	chr14:20472000-20476500	CAAGCTGCCAAAACCTTCCA	CCCTGTATGACCTCAACTGACA
RELA	chr11:65425500-65430000	CCTTTCTGCACCTTGTCACAC	ATCTGCCGAGTGAACCGAAA
RPS26	chr12:56041893-56044223	TCCCAATTCACTCGGTTCTCA	TCATGGCAGAAGGGGAAACA
RPS3	chr11:75400000-75404500	CTGTTCAGGTCAGCAGTGTAC	CAATGGGCAGAACCATGTCA
SLC20A1	chr2:112645500-112650000	CACCGTCTCACTTTCCCTCA	GAGGGGTCTCAAGTCAGCATAA
SLC2A1	chr1:42932000-42936500	GTCCTTGTACCTCATCCACTCA	TTTGGGCTTTCTGTCCCTCA
TMF1	chr3:69020000-69024500	TGAAGGAAGTGGCCTTGCTA	CGAGGCAAGAAATGCACGTA
TPBG	chr6:82364244-82367417	TCTGTCTGTCCCTATGGCCTA	TTCCCCAGAAGGCACTTCAA
TRA2A	chr7:23515500-23520000	ACAAGCAAGGCAAGCACTAC	CTCCCCAGCTTTAAAGACATCATAC
TUBA1A	chr12:49184796-49189324	TGGCTCTCTCTGCATGGTTTA	GGGACGTGGTAGGAACTCAATA
UBC	chr12:124911704-124914601	TGTCAGATGCAACCGAGGAA	GCCACGCATGGCTATTGAAA
UGP2	chr2:63870000-63874500	CCAGTGGATGGATGAGGAGTAA	TGAGTACCAGCTCCAAAGCA
UQCRC2	chr16:21965500-21970000	AGTTGCCCACAACCCTTCA	AGGAAGGCCATGTGACAACA
VDAC2	chr10:75218000-75222500	CTGGTGAAAGGTGACTGGATCA	ATCGTGAGGACAGCACCAAA
VEGFA	chr6:43775500-43780000	CTCTGCTGTGTCTGCATCAA	GCAGCTAGGCGCAAAGTATA

Table 4. Information on primers used in qPCR (genotypoing).

SNP	STA (Specific Target Amplification) primer	LSP (Locus-Specific Primer)	ASP1 (Allele-Specific Primer 1)	ASP2 (Allele-Specific Primer 2)
MDH2 (p.E300K)	CTCCTCAAAAGAGGAGACTTTGC	ACTTCTCCACACCGCTGCT	CCGATGCCCAGGTTCTTCTC	CCGATGCCCAGGTTCTTCTT
COA3 (p.S78X)	CCTCGTCTTCTAGCTCATCTAGG	GGCCCTGGTGTTGGCTATTT	GAAACGCTCCTGGGAAATCG	GGAAACGCTCCTGGGAAATCT
KRAS (p.G12D)	CTGAATTAGCTGTATCGTCAAGGC	CCCAGGTGCGGGAGAGA	CACTCTTGCCTACGCCAC	GCACTCTTGCCTACGCCAT
BUB3 (p.P210L)	CTGTGACATTTGAAGGCATACTTCT	TCTATTGAAGGCCGAGTGGCA	TTCTTCTGTACCTCAGGGCTTG	TCTTCTTCTGTACCTCAGGGCTTA
C12orf65 (p.S89X)	GTTCTGATCAACTGATCTTGTCTGA	GCAACTGCGTGGTGCTGAA	GGCACTTTACAACGATGCCTG	GGCACTTTACAACGATGCCTC
HSPA5 (p.E308Q)	GGCCAAACGGGCCCT	CCCGAGTCAGGGTCTCAGAA	GTCTTCTCAGCATCAAGCAAGAATTG	GTCTTCTCAGCATCAAGCAAGAATTC
CNDP2 (p.M317I)	GCCTTCCCTGGTCAAGTC	CAGTCACCCTCATTACCTGGCT	GGCTCAACACCAAAAACTGTCTTC	GGCTCAACACCAAAAACTGTCTTA
ZNF638 (p.A1703S)	GGAGAAGTGGAAGAGCTACCT	CCCTTACAGTATCTCCTTCATTTCCTTTAGT	TTGAATGAGTCAGCAGACATAACTTTTG	TTTGAATGAGTCAGCAGACATAACTTTTT
TMEM98 (p.E81K)	GTGGGACATGAGACCCGA	TCGTTATCACCAACCCCCACA	AGGCATCTTCGATCCAGTCTTC	GAGGCATCTTCGATCCAGTCTTT
C3orf37 (p.E60Q)	AATCCAACAGCCCAGTGC	GCATGGGAGCAATGATACGCT	CTTCTGTCTCGACTGCACTTTG	CTTCTGTCTCGACTGCACTTTC
CCDC88A (p.I970T)	TGATGACAGGTATAAACTTTTGGAATCA	CGAGCTTCTAAAGCAGCAATTTTTTCTTCT	TTAGAATCCACTCTTAAGAAGTCTCTTGAAAT	AGAATCCACTCTTAAGAAGTCTCTTGAAAC
G3BP1 (p.R429W)	TGCCAGGGAAGGCGAC	TCTCATTCCACCACCCAGCC	ACCGACGAGATAATCGCCTTC	GACCGACGAGATAATCGCCTTT
PNPT1 (p.L763F)	GTGCTTCAGTCGCCAGC	GGTTCTCCCATTACAATACTACTTCTGTCA	CTACAACCGTGGTCAGAACTTTG	CTACAACCGTGGTCAGAACTTTC
EBAG9 (p.S62F)	ACAGTTGATTATTCATCAGTTCCTAAGC	GGTGGGTGCATCTTCATCCC	AGACAGATGTTGAAGAGTGGACTTC	AGACAGATGTTGAAGAGTGGACTTT
RAD21 (p.E415G)	GCTGGTCCTCTCTAGGAACC	ACCGCTTGTACCAGAAGACCTT	CCTCTGGATTTTCAAATTCTTTGAGGAATT	CCTCTGGATTTTCAAATTCTTTGAGGAATC
RAD21 (p.E415K)	GCTGGTCCTCTCTAGGAACC	ACCGCTTGTACCAGAAGACCTT	CCTCTGGATTTTCAAATTCTTTGAGGAATTC	CCTCTGGATTTTCAAATTCTTTGAGGAATTT
AP5M1 (p.G145D)	GGAGTTTCACAAGGCTTTGAATTT	CAAGTCAGGCAACTGGCTCA	TTTTTGGGATACAGGATTTTCTTTATTCAGG	CTTTTTGGGATACAGGATTTTCTTTATTCAGA
REV1 (p.D1201H)	CAATGGAAGAAGACATTCTCCAAGT	CCGATTGCTGCATCAGCCTTT	TTGTGAAATACTGTACTGATCTAATAGAAGAAAAAG	TTGTGAAATACTGTACTGATCTAATAGAAGAAAAAC
FTSJD1 (p.K38X)	GCCCAGATATTCTTGCTGACAT	CTGGGATCTGGTAACTGCCACT	TTTGCCAAGAACTTTTCTTATGGCA	CTTTGCCAAGAACTTTTCTTATGGCT
MAN2A1 (p.E967Q)	GTTATCATGGATCGAAGACTCATGC	GCTGTAATCTTGTTATCCTGGATACCTTGC	GCAAGATGATAATCGTGGCCTTG	GCAAGATGATAATCGTGGCCTTC
NDE1 (p.G313C)	GGACAACCAGCGGCAA	GAGAGACGGCCAAGCAGC	CGTGTCCAACCCCTTATCACC	CGTGTCCAACCCCTTATCACA
HSDL2 (p.P138A)	TGGACTGATATTGAGGATATGAGCA	CCAGAGGCACCTACCTTGCA	GAGCAACTTTGCTCTTTTTCAAATAAGG	GAGCAACTTTGCTCTTTTTCAAATAAGC
CASP3 (p.E84K)	CACGCATCAATTCCACAATTTCTT	GCAGCAAACCTCAGGGAAACA	CACGTGTAAGATCATTTTTATTCCTGACTTC	CACGTGTAAGATCATTTTTATTCCTGACTTT
CABIN1 (p.R244L)	GCTCCTTCTCCCGCACAA	GCAGCTGAGACACAGGCGAT	TCAGCGCTTGCCTCTTTTTTC	CAATCAGCGCTTGCCTCTTTTTTA
WDR3 (p.E770Q)	AGCAGCTGAGAGGATTATGGA	GGAACCTCTTTCCCTGCAGC	AGGCTATTGAGTTGTACCGAGAAG	AGGCTATTGAGTTGTACCGAGAAC
CASP3 (p.R149T)	AAATGGACCTGTTGACCTGAAAA	ACCACGGCAGGCCTGAA	ACTTTTTCAGAGGGGATCGTTGTAG	ACTTTTTCAGAGGGGATCGTTGTAC
RAD21 (p.D464H)	CCAGCTTTTCGCTTAACTCCC	TGATGGAGGCCAGCAGAACA	TGGAGGCATAGCTGACTCATC	TGGAGGCATAGCTGACTCATG
RCN2 (p.S284C)	CACTGGTGAGAAACAAGTCCG	GCACAAGAGGAGGCGCTTCA	GGGTTTTCCAGAATCTCTTCTTCAG	GGGTTTTCCAGAATCTCTTCTTCAC
CUTC (p.V258L)	CACCAGGATGTTCTTTGCGA	TGGGAGCCTCACTTTCTTGCT	GCATTCAAAGTCCTTACTTTGGTCAC	GCATTCAAAGTCCTTACTTTGGTCAA
IDH3A (p.V218L)	CATCATGCGGATGTCAGATGG	TACATCTCATTAAATTTAATATCTTTACAGCTTTCTGCAA	GCTTTTTCTACAAAAATGCAGGGAAG	GCTTTTTCTACAAAAATGCAGGGAAC
RNF219 (p.S401C)	CAGCTCCTTGTACTCCTTTGTC	AACCTCCTGCTTGGACCACA	CCCTTAGTTGCCTTCAGCTCA	CCCTTAGTTGCCTTCAGCTCT
MAP4K4 (p.S579I)	CTCTGCCTGTCTGCTTAGACT	ACCGAGCGCGAGAGGTG	TGGGCTTCAGGGGAGC	ACTGGGCTTCAGGGGAGA
CASP3 (p.E123K)	GCAAAAGGAGCAGTTTTGTTTGT	AGGTCAACAGGTCCATTTGTTCCA	GTGCTTCTGAGCCATGGTG	TGTGCTTCTGAGCCATGGTA
GAPVD1 (p.E1214Q)	CGACAACGAGTGAGATAAGCAA	CGCTGTGTGTGCCGTTTTG	GGGGCTCTTTTTCTGTAGTCCTC	GGGGCTCTTTTTCTGTAGTCCTG
JMJD1C (p.P1011L)	CCTGGAAAGAAAGGAAGGCA	CCACCAGCATTTACTGGCATCA	CAGCTATAGTAGTCTTTCCCCTCC	GCAGCTATAGTAGTCTTTCCCCTCT
BAZ1A (p.R859P)	CTTCAATAAGCTGGTCTAGCTGT	CAGCTGCCAAAACCAGTGCA	CACAAGAACTGTAAAAGCACCACC	CACAAGAACTGTAAAAGCACCACG
PHTF2 (p.S204R)	GAAAGCAGCATGCCAGAGAT	GGAGCAGTTCAGAACCACGG	TCTCTGAAGACAGTGCCAACG	GATCTCTGAAGACAGTGCCAACT

MUC4 (p.S2055F)	TTAGTGACAGGAAGAGGCGT	CACAGGTCAGGACACCCCT	GTGTCACCTGTGGATACTGAGG	GTGTCACCTGTGGATACTGAGA
USP32 (p.S67C)	ACTATTAAATTATTGAAGTGCAGCCCT	GAGTGCCTCCAAAGGTTGCT	CCCTTTGGATGTTCCACCAAAAG	CCCTTTGGATGTTCCACCAAAAC
NUDT3 (p.E123Q)	CTGCACGGGTTTGTGATACT	TGCTGGAAGACTGGGAAGATTCA	GCAGCACTTTTATGGCGTCTTC	GCAGCACTTTTATGGCGTCTTG
TMEM8A (p.K761R)	AGTGTCACGTCACTGCGTA	GCCTGCTCGCAGAAATTCCC	CAGTTCCTCCCGATCGTTCT	CAGTTCCTCCCGATCGTTCC
MGST2 (p.P44T)	CTCCACACAGTTTTGTTGTGC	CAAAGTTACGCCCCCAGCA	CCCGAAATACTCTCTCAAACTCTGG	CCCGAAATACTCTCTCAAACTCTGT
KIAA0430 (p.D427N)	GTAACCGTTGCCCACATCAA	CAAATCTGCGGAGACTCTGCC	TGCAAAGAATGCCGCTGATG	CTGCAAAGAATGCCGCTGATA
ROCK1 (p.D1158H)	AACGTGAAACAGTTTATCTATGTCCAAT	AAACAGTATGTTGTGGTAAGCAGCAAAA	CATAGATGGATTGGATTGCTCCTTATC	CATAGATGGATTGGATTGCTCCTTATG
PAEP (p.M19I)	CGTGGCCCTGGTCTGT	CCCTGCCAACTTTGGGAGC	TGGTGTCCCGGCCATG	TGGTGTCCCGGCCATT
CNTRL (p.E1329Q)	CATCCACCTTTCTTCCCGC	GCAACGTCCCTGAACACCAT	CTTCTTTGATTTTAAATGCTGCATTATGTCTTC	CTTCTTTGATTTTAAATGCTGCATTATGTCTTG
GPR82 (p.L275V)	AAGCAAGACAGGTGAGAATGTTT	GTTTGCTTCCTTCCTTATAGTATTTTTAAACCCA	TGCTGACAGTTATCTCTTTGGTGTAG	TGCTGACAGTTATCTCTTTGGTGTAC

Table 5. Information on primers used in ddPCR.

chr	position	ref_allele	alt_allele	gene	Forward primer	Reverse primer	WT probe	Mutant probe
chr7	75695609	G	A	MDH2	GGGTTTCTCTAACAAGCACTTTC	AAGAGGAGACTTTGCCGATG	TTCTTCTCGATGCCC (AntiSense)	TTCTTCTCGATGCCC (AntiSense)
chr12	25398284	G	T	KRAS	commecially available in Bio-Rad (PrimePCR™ ddPCR™ Mutation Detection Assay Kit: KRAS WT for p.G12D, and KRAS p.G12D, Human #186-3112)			
chr10	124921804	C	T	BUB3	TTCTCTCTTGGCAGGGTTATG	GGCATACTTCTTCTTCTGTACCT	ATTTGGACCCAAGCC (Sense)	ATTTGGACCCAAGCC (Sense)
chr2	55863435	C	G	PNPT1	ACTGTGAAATAGGTTCTCCCATTA	CTTCAGTCGCCAGCTACAA	TCTGTCATTCAAAGTTCTGA (Sense)	TCTGTCATTGAAAGTTCTGA (Sense)
chr8	110566980	C	T	EBAG9	CCCTTTGTGGCTGATAATGTTG	CCATTCCCTCCTTCGATCTTT	ATCCCAGGAAGTCCAC (AntiSense)	ATCCCAGGAAGTCCAC (AntiSense)
chr5	109183414	G	C	MAN2A1	GTGTTAGGTCAGATTGAAGTTATC	GATTAGCTGTAATCTTGTTATCCTG	ACCTTGCTCAAGGCC (AntiSense)	ACCTTGCTCAAGGCC (AntiSense)
chr1	118496669	G	C	WDR3	CTGTAGGCTGAGAGGATTATGG	CCCTGCAGCTTTACAAATGG	CGAGAAGAAACTGCAA (Sense)	CGAGAAGAAACTGCAA (Sense)
chr15	77241460	C	G	RCN2	TTTGAAGGCGCTTCATCTAATTG	ATGGAGCTGTCTGCCATAATC	AAGCTCTCTGAAGAAGA (Sense)	AAGCTCTCTGAAGAAGA (Sense)
chr10	101515446	G	T	CUTC	CATGGGAGCCTCACTTTCTT	GGCTACACCAGGATGTTCTTT	AACAGATGTGACCAA (Sense)	AACAGATTTGACCAA (Sense)
chr15	78455889	G	C	IDH3A	AAATACAGGCGGATGTCAGAT	CAAACATACTGTATCAAGGTACATCTC	TCTGCAACTTCCCT (AntiSense)	TCTGCAACTTCCCT (AntiSense)
chr2	102477318	G	T	MAP4K4	CTCTGATGCAGGTGGAAGATAG	CACTGGTGGCTCCAATACTC	AACCACAGCTCC (Sense)	AACCACAGCTCC (Sense)
chr4	140599768	C	A	MGST2	GCAAGTTGGAAAGGCAAGATTA	GATCCATGAAGGGTGGAAGAA	ACTCTGGTGACCC (AntiSense)	ACTCTGGTGACCC (AntiSense)
chrX	41587102	C	G	GPR82	GTCATCCAGATTCTAC	ATTGCTGACAGTTATC	ATGTTCTACACCAAAG (Sense)	ATGTTCTACACCAAAG (Sense)

RESULTS (PART-II)

Translation of single cell expression signatures into therapeutics for a metastatic renal cell carcionoma patient

1. Enrichment of subclonal cancer cells in paired lung metastasis

Paired primary and lung metastatic RCC surgical samples were harvested from a patient (Figure 1B). This 43-year-old male patient was initially diagnosed as localized clear cell RCC and had radial nephrectomy. Unfortunately, metachronous lung metastasis was detected 1 year after the nephrectomy, which was surgically removed also. In spite of following pazopanib, everolimus, and high dose interleukin-2 chemotherapy, the patient showed intrinsic resistance to the palliative therapies disseminating cancer cells to other organs including the bone, lung, pleura, and brain. Parts of the surgical samples were minced into ~1mm^3 blocks and then transplanted into the subrenal capsule of immune-deficient mice to enrich cancer cells preserving molecular and functional heterogeneity (Figure 1B).

Genomic signatures of pRCC and mRCC were profiles by integrated analysis on WES and aCGH. Somatic single-nucleotide variants (SSNVs) commonly found (n=115) including exon 2 missense mutation (D121G) of the VHL gene in disease progression maintained with higher level of VAF compared to SSNVs found exclusively in pRCC (n=180) and mRCC (n=192) (Figure 20A). Copy number profiles between pRCC and mRCC showed analogous pattern with a few differential copy number status (Figure 20B). When it comes to the somatic mutations that are significantly identified in ccRCC by TCGA profiling (69), pRCC and mRCC showed shared SSNVs and SCNAs only except in 5q amplification that was not detected in mRCC. Next, we assessed propagation of cancer cells derived from patient tumors to the xenograft model. Similar histopathological and cellular morphologies were observed

with enriched through xenograft (Figure 21A and B). SSNVs (n=59) commonly preserved from patient pRCC to PDX mRCC showed higher cellular prevalence (Figure 21C and D). Moreover, significant copy number variation events were also preserved in PDX tumors (Figure 21E and F). Together with these genomic signatures, we inferred evolutionary pattern between pRCC and mRCC. By implementing a PyClone algorithm (65), cellular prevalence of somatic mutations and the number of shared subclones between pRCC and mRCC were estimated (Figure 22. For the detailed, see Figure 23). Subclones tended to be enriched in mRCC with higher cellular prevalence. Especially, fluctuation of cellular prevalence was greater in subclones with lower cellular prevalence (Subclones 4 and 5). Taken together, these data demonstrated that driver mutations shared across pRCC and mRCC were occurred in early and played a pivotal role in disease progression.

To decipher ITH of transciptome, we profiled cancer cells (n=46 and 36 for the primary and metastatic tumor, respectively) recovered from the xenograft tumors by single cell RNA-seq. Robust investigation for quality control was applied in single cell samples to exclude samples with low quality (Figure 24). Compared with the gene expression of normal kidney cortex, the cells showed low expression of stromal cell-related genes and high expressions of genes that are related with worse prognosis of RCC (Figure 25). These characteristics show concordance with genomic profiling (Figure 21) and indicate the cells were rarely contaminated by mouse stromal cells.

Gene expression of cancer cells from the xenograft tumor of the lung metastatic RCC (mRCC cells) were different with those of the primary RCC

(pRCC cells) (Figure 26). In the spatiotemporal vector virtually created by Principal Component Analysis (PCA), pRCC cells were positioned separately with mRCC cells (Figure 26A). Moreover, gene expressions of pRCC cells were more homogeneous compared with those of mRCC cells. Although the mRCC cells showed more heterogeneous gene expression, expression of genes related with epithelial to mesenchymal transition (EMT) (Figure 26B) and worse prognosis of RCC (Figure 26C) were significantly up-regulated in those cells.

2. Prediction of effective drug selection from single cell transcriptome profiles

Based on the gene expression signatures of the pRCC and mRCC cells, activated oncogenic signaling pathways were predicted. Compared with the pRCC cells, the mRCC cells showed significant over-expression of EGFR, Src, Raf, and MEK signaling pathway-related genes (Figure 27B and 29). In contrast, VEGFR, c-Met, PI3K/Akt, and PDGFR signaling pathways were predicted to be activated in the pRCC cells (Figure 27B). Those predictions were well correlated with *in vitro* drug screening results (Figure 27A and 28). Targeting agents against EGFR (Gefitinib, Erlotinib, and Afatinib), Src (Dasatinib), and MEK (Selumetinib) showed much lower IC50 in the mRCC cells whereas the pRCC cells were more sensitive to c-Met (Tivatinib, Foretinib, and Crizotinib) and PI3K (BKM120)-targeting agents (Figure 28 and 29). Surprisingly, although the patient was treated with Everolimus and

Pazopanib in the clinic, mTOR and VEGFR signaling pathways were not indicated to be activated in both the pRCC and mRCC cells (Figure 27B). Unfortunately, the prediction was realized in not only experimental *in vitro* drug screening (Figure 27 and 29) but also clinical treatment results.

The *in vitro* drug screening results for Afatinib and Dasatinib were validated further since they showed the least IC50 in the mRCC cells (Figure 28 and 29). In a 3D cell culture system that simulated real tumor tissue more accurately (Figure 30A) and *in vivo* patient-derived cell xenograft (PDX) animal models (Figure 30C), Afatinib and Dasatinib showed significant treatment effects on the mRCC cells. Specific on-target effects of Afatinib and Dasatinib against the EGFR and Src signaling pathway, respectively, were confirmed *in vitro* (Figure 30B) and in vivo (Figure 30E). Both Afatinib and Dasatinib significantly decreased the number of proliferating cancer cells and increased apoptotic cell number in the PDX models (Figure 30E).

Single cell RNA-Seq data further indicated that EGFR and Src signaling pathway might be activated in the different populations of mRCC cells (Figure 27B). Based on the heterogeneity, combination treatment using EGFR and Src targeting agents would make therapeutic effects on broader cancer population and have better treatment results (Figure 31A). Based on this speculation, combination therapy using Afatinib and Dasatinib was compared with Afatinib or Dasatinib single treatment. The combination specifically inhibited both EGFR and Src signaling (Figure 31B) and showed significantly better treatment effects on mRCC cells than the single therapies in 2D (Figure 31C) and 3D (Figure 31D) in vitro culture system. Furthermore, we

successfully confirmed such treatment effects in the mouse model. (Figure 31E). Taken together, these data demonstrate that dissecting ITH of transcriptome at single-cell resolution could enlighten us to predict drug treatment responses and select the most effective drugs.

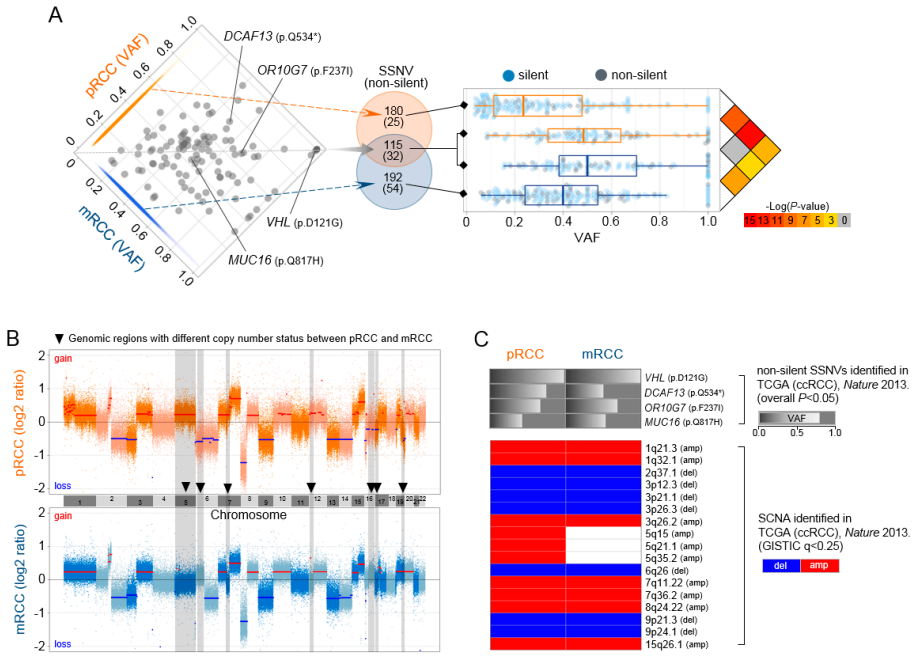


Figure 20. Comparative mutation profiles between pRCC and mRCC tumors in PDX models. (A) Variant allele frequencies (VAF) observed across pRCC and mRCC WES data. Left, a scatter plot represents distribution of VAF for SSNVs exclusive in pRCC (top, orange) and mRCC (bottom, blue) and shared in pRCC and mRCC (middle, gray). Center, the numbers of exclusive and shared SSNVs are shown in Venn diagrams. Right, boxplots demonstrate mutual significant differences of VAF distribution only except between pRCC and mRCC for shared SSNVs. Two-tailed t-test was applied to test statistical significance of P-value. (B) Copy number profiles from aCGH data. Break points of CGH probes were detected by using the CBS (circular binary segmentation) algorithm. Different regions of copy number between pRCC (orange) and mRCC (blue) are highlighted in shadow with arrows along the chromosomes. CBS-derived copy gains and losses were defined when log2 ratios were greater positively or negatively than absolute 0.25, respectively. (C) Overlapping of detected mutations with common ccRCC features identified in TCGA for ccRCC (69). Each filtering criteria with statistical significance for SSNV and SCNA is denoted, respectively. Non-silent SSNVs annotated in our data were only considered to overlap with TCGA data.

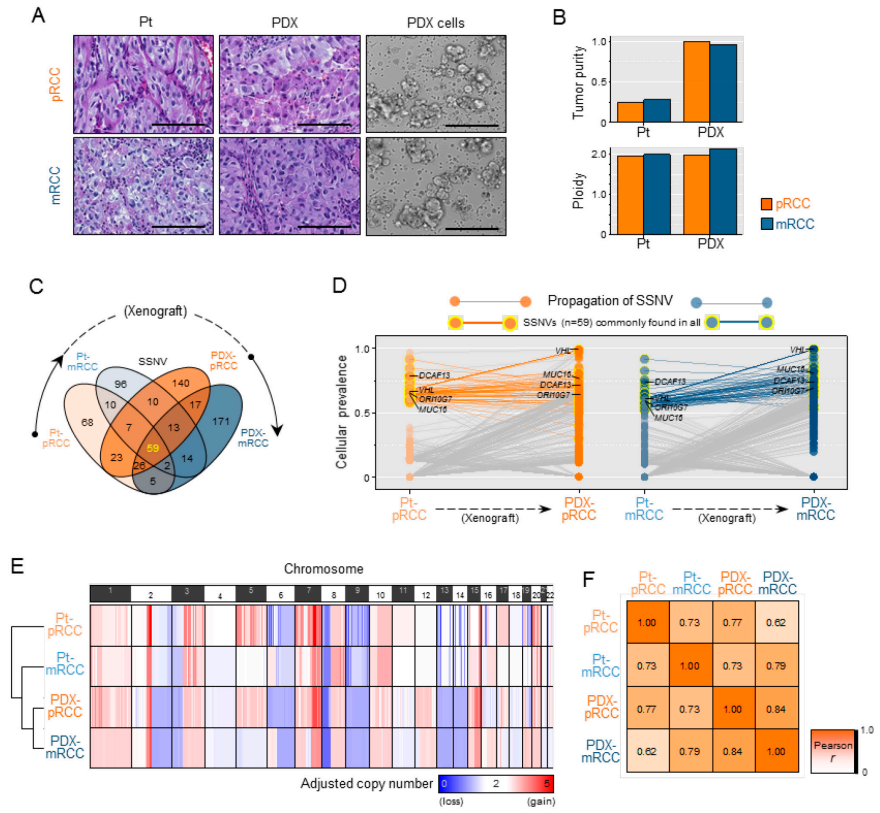


Figure 21. Identification of retaining driver mutations through xenograft propagation. (A) Morphological similarity between parental tumors and their xenografts by H&E staining and bright-field observation for tumor tissues and PDX cells, respectively. (B) Computational estimation of tumor purity and ploidy by implementation of the ABSOLUTE algorithm with WES and aCGH data. (C) Venn diagrams of SSNVs across four tumors labeled. The number of intersection for all sets is colored in yellow. (D) Tracking cellular prevalence changes of SSNVs in the process of xenograft. SSNVs (n=59) maintained highly in all samples are highlighted with yellow outline circles and thicker orange extension lines. Non-silent SSNVs significantly observed in ccRCC (Figure 20C) are denoted. (E) SCNA along the autosomal chromosomes that were adjusted with estimated tumor purity and ploidy. The order of samples was determined by average linkage clustering in Euclidean distance similarity metrics. A heatmap demonstrates the overall concordance of SCNA that are significantly observed in ccRCC (Figure 20C). (F) Pearson correlation coefficient (*r*) of copy number profiles between samples.

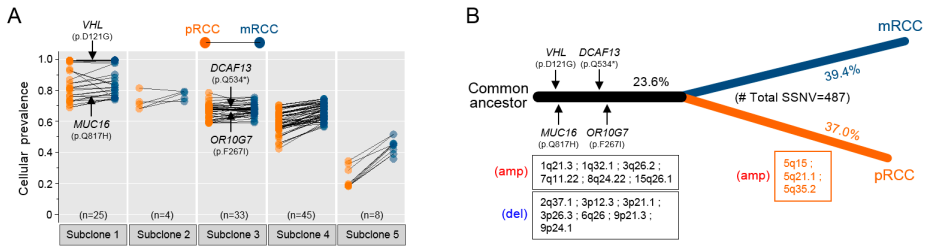


Figure 22. Cellular prevalence of shared subclones and inferred tumor evolution between pRCC and mRCC. (A) Identified five subclusters commonly found in pRCC and mRCC by integrated analysis of aCGH and WES data. The mean cellular prevalence of each SSNV was estimated by implementing a PyClone algorithm. Dominant subclones harboring SSNVs with higher cellular prevalence are ordered from the left. (B) Inferred phylogenetic evolution between pRCC and mRCC. Branch and trunk lengths are proportional to the number of SSNVs. SSNVs and SCNAs that are significantly observed in ccRCC TCGA data are denoted. See also Figure 23.

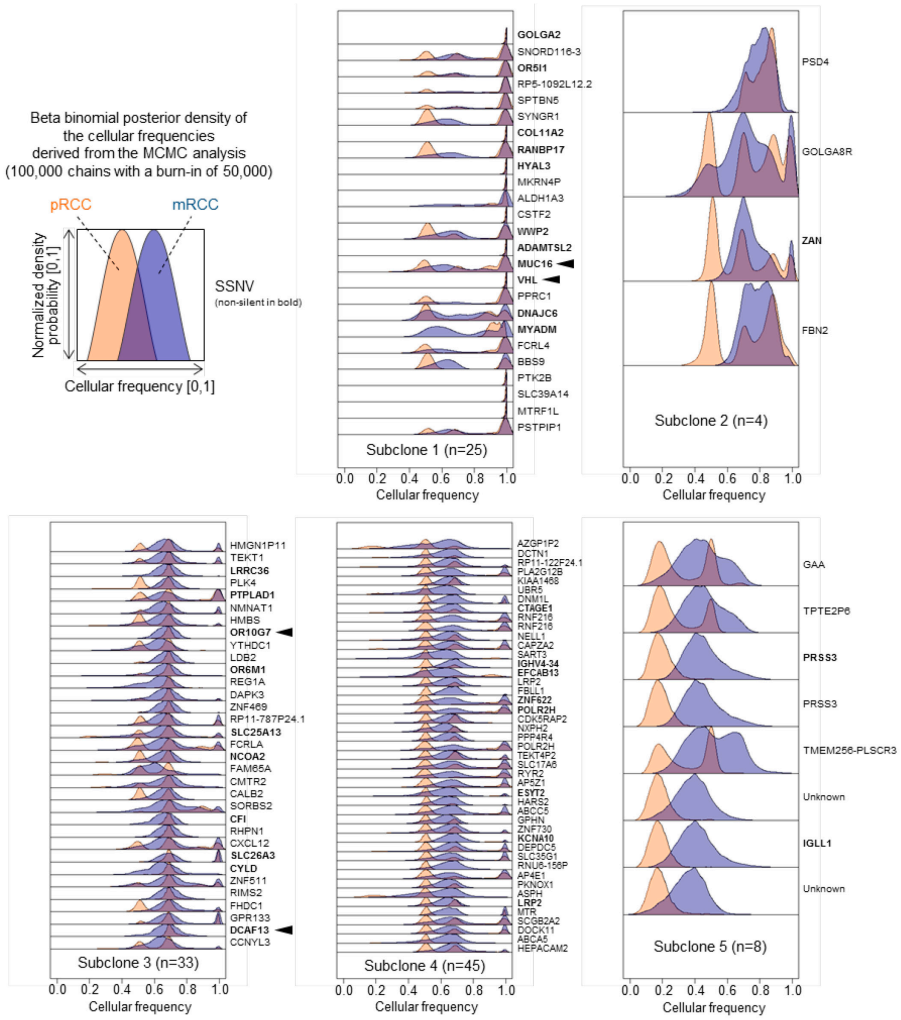


Figure 23. Detailed information for estimated cellular frequencies relevant to each SNV between paired primary and metastasis. Gaussian kernel density estimates were applied in a post-burn-in MCMC traces in the PyClone algorithm. Non-silent mutations are highlighted in bold. Arrows indicate underlying driver mutations in Figure 20.

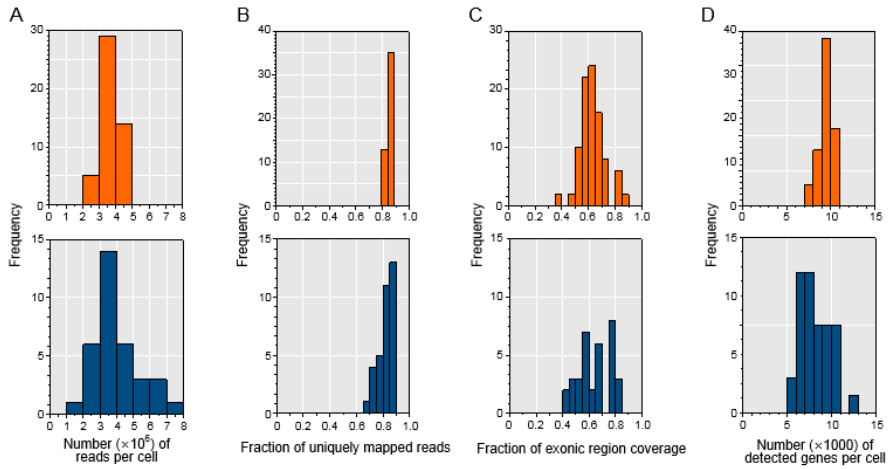


Figure 24. Performance assessment of single-cell RNA-seq data. (A-D) Histograms of single cell frequencies in the number of generated RNA-seq reads per cell (A), uniquely mapping rates (B), exonic regional coverage rate (C) and the detected number of genes per cell (D). By filtering criteria of > 1M reads per cell, > 60% unique mapping rate, > 35% exonic region coverage and > 5,000 detected genes, three single cell samples were not included in this study.

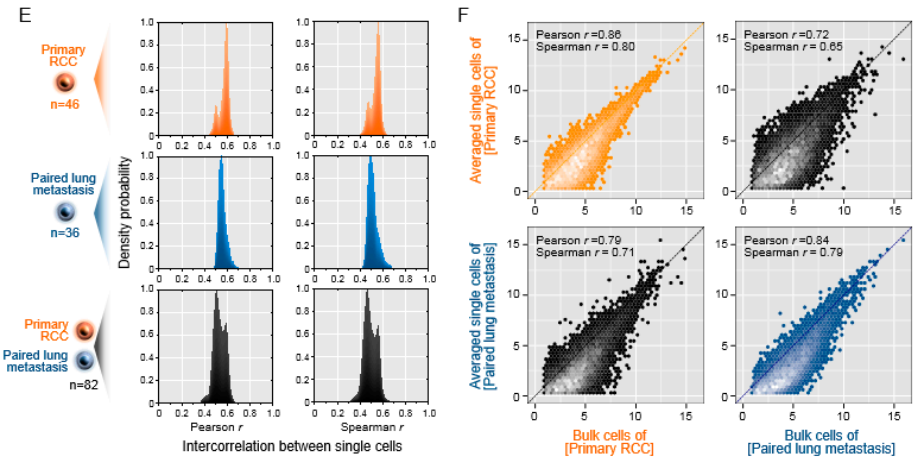


Figure 24. Performance assessment of single-cell RNA-seq data. (E) Cell-to-cell variation of gene expression within (top and middle) and between (bottom) pRCC and mRCC single cells. Inter-correlation between single cells was estimated as Pearson and Spearman correlation coefficients. (F) Scatter plots show similarities of gene expression between averaged single cells and bulk cell population when the pair is matched (orange or blue) and unmatched (gray). Black dotted line is the x=y line with correlation coefficients (Pearson and Spearman r) for linear fit.

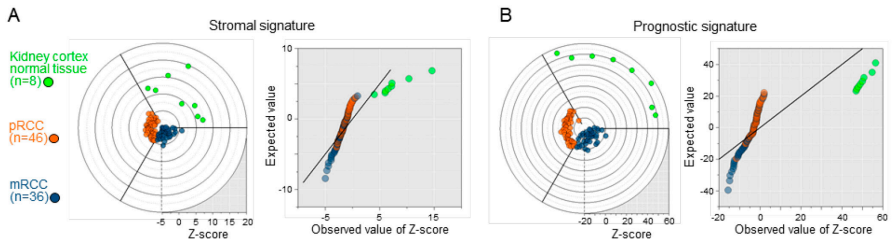


Figure 25. Expression signatures of tumor cells compared to normal tissues. (A-B) To compare and normalize expression profiles of our datasets, normal kidney expression profiles were downloaded from the GTEx portal (n=8, Ver.3) and used in this study as converted TPM values. To identify outlier values in normal distribution, Z-scores were estimated for gene sets of the stromal signature (A) and the ccRCC prognostic signature (B). Outliers were observed both in radial and Q-Q plots.

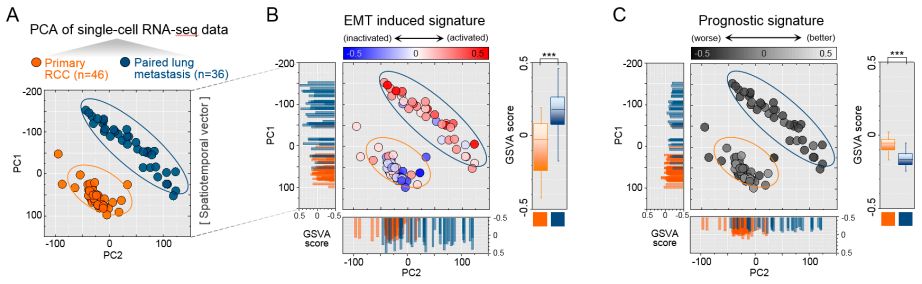


Figure 26. Recapitulation of metastatic expression signatures at single-cell resolution. (A) Global differences of expression profiles between pRCC and mRCC determined by principal component analysis on single-cell RNA-seq data. Individual tumor cells are positioned as dots (orange=pRCC, blue=mRCC) in the PC1-PC2 plane and ellipses represent 95% confidence around each group. (B-C) Gene set activation analysis on the EMT induced signature (B) and the ccRCC prognostic signature (C). Positions of each dot and ellipse were fixed as in (C) and colors represent the relative status of expression signature across single cells. Boxplots demonstrate significant differences of expression signatures between pRCC and mRCC single cells. *** P<0.001, two-tailed Student t-test.

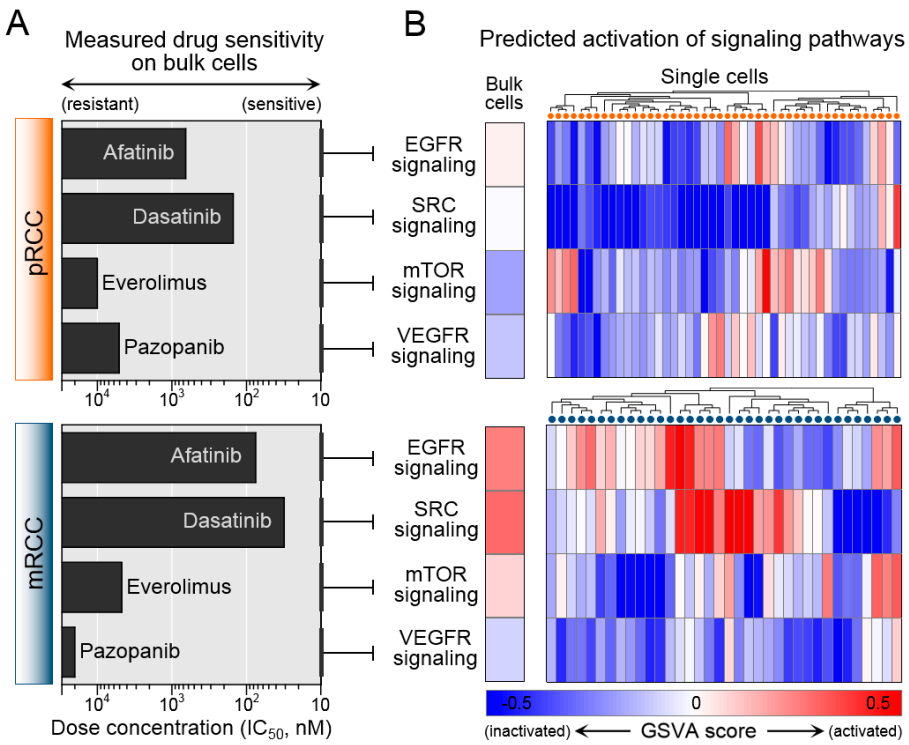


Figure 27. Identification of activated signaling pathways that are sensitive to anti-cancer drugs. (A) Measured drug sensitivity (IC_{50}) of Afatinib (targeting EGFR signaling), Dasatinib (Src signaling), Everolimus (mTOR signaling) and Pazopanib (VEGFR signaling). (B) Predicted activation of signaling pathways of the given gene sets; EGFR signaling, [Reactome] Signaling by constitutively active EGFR; Src signaling, [Extracted from (70)]; mTOR signaling, [Reactome] mTOR signaling; VEGFR signaling, [PID] Signaling events mediated by VEGFR1 and VEGFR2.

Drug	Target	Primary RCC	Paired lung metastasis
Gefitinib	EGFR	6400	290
Erlotinib	EGFR	>20000	640
Afatinib†	EGFR/HER2	650	25
Tivantinib	C-Met	520	13000
Foretinib	C-Met/VEGFR2	200	860
Crizotinib	C-Met/ALK	1200	3600
Selumetinib	MEK	2400	210
Vemufafenib	B-Raf (V600E)	2800	7700
Temsirolimus	mTOR	9500	10000
Everolimus*	mTOR	10000	11000
BKM120	PI3K	500	930
Cabozatinib	VEGFR2	3800	8700
Vandetanib	VEGFR2	1100	1700
Sunitinib	VEGFR2/PDGFRβ	2200	2500
Sorafenib	Raf-1/B-Raf/VEGFR2/PDGFRβ	4500	8500
Pazopanib*	VEGFR1-3/PDGFR/FGFR/c-Kit/c-Fms	5100	> 20000
Nintedanib	VEGFR1-3/FGFR1-3/PDGFRα/β	1000	2400
Dovitinib	FLT3/c-Kit/FGFR1/3/VEGFR1-4	1200	400
Dasatinib†	Src, Abl	190	31

Figure 28. The results of drug screening in pRCC and mRCC. Afatinib and Dasatinib were selected as effective drugs (denoted as †) compared to Everolimus and Pazopanib (denotes as *) that were clinically failure to drug efficacy.

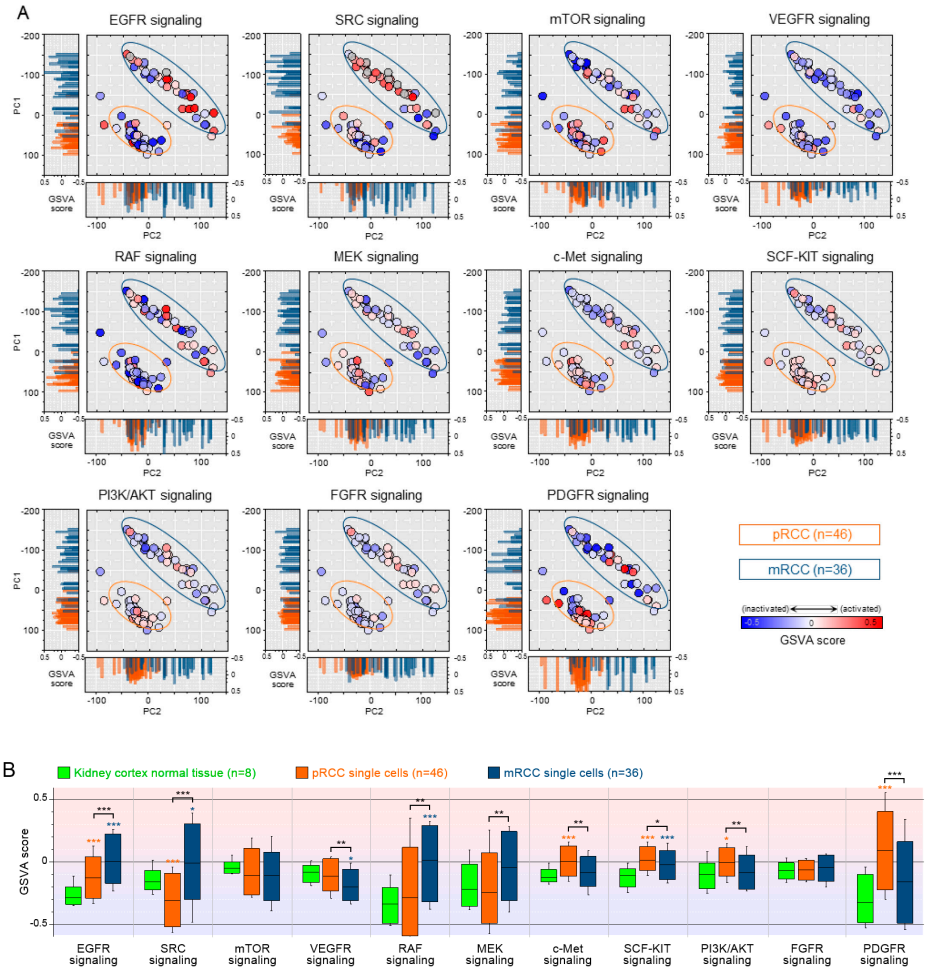


Figure 29. Identification of activated signaling pathways that are targeted by anti-cancer drugs. (A) Heterogeneous activated status of the given targetable signaling pathways at single cell resolution. Positions of each dot and ellipse were derived from PCA and fixed as in (Figure 26A) and colors of dots represent the relative status of expression signature across single cells. (B) Boxplots demonstrate significant differences of expression signatures compared to normal kidney tissues or between pRCC and mRCC single cells. * P<0.05, ** P<0.01, *** P<0.001, two-tailed Student t-test.

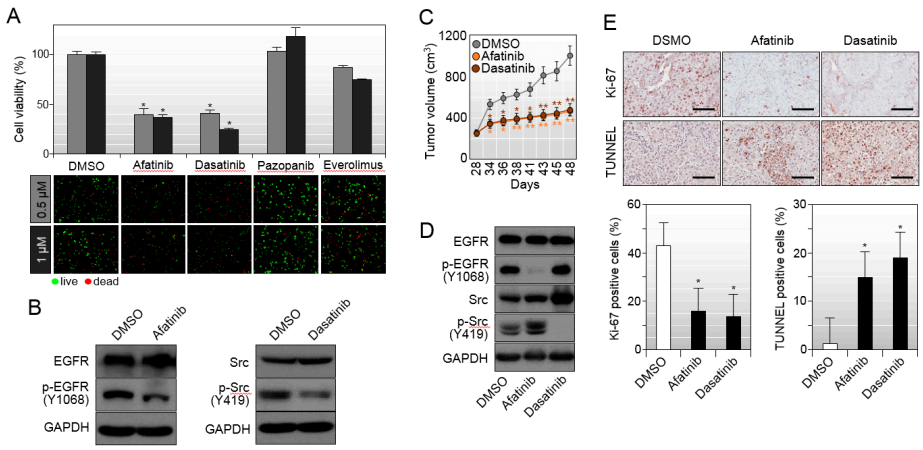


Figure 30. Validation of drug efficacy in *in vitro* 2D and 3D and *in vivo* models. (A) Measured cell viability in a 3D cell culture model. The fraction of live cells were estimated from counting green fluorescent dots by the total number (green plus red fluorescence) of cells in a selected region. Standard deviation was calculated from triplicate data. Student t-test was applied to validate statistical significant difference to mock-treated samples (DMSO) each n=3, * $P< 0.05$. (B) Relative gene expression of RT-PCR upon drug treatments. (C-E) Application of drug treated responses in *in vivo* mouse models. (C) Measured tumor volume in mouse xenografts. Student t-test was applied to validate statistical significant difference to mock-treated samples (DMSO) each n=5, * $P< 0.05$. (D) Western-blot validation. GAPDH was used for a housekeeping marker in (B) and (D). (E) Measurement of proliferation and apoptosis status by Ki-67 and TUNEL staining assays. Student t-test was applied to validate statistical significant difference to mock-treated samples (DMSO) each n=3, * $P< 0.05$. This experiment was performed by the Institute for Refractory Cancer Research, Samsung Medical Center.

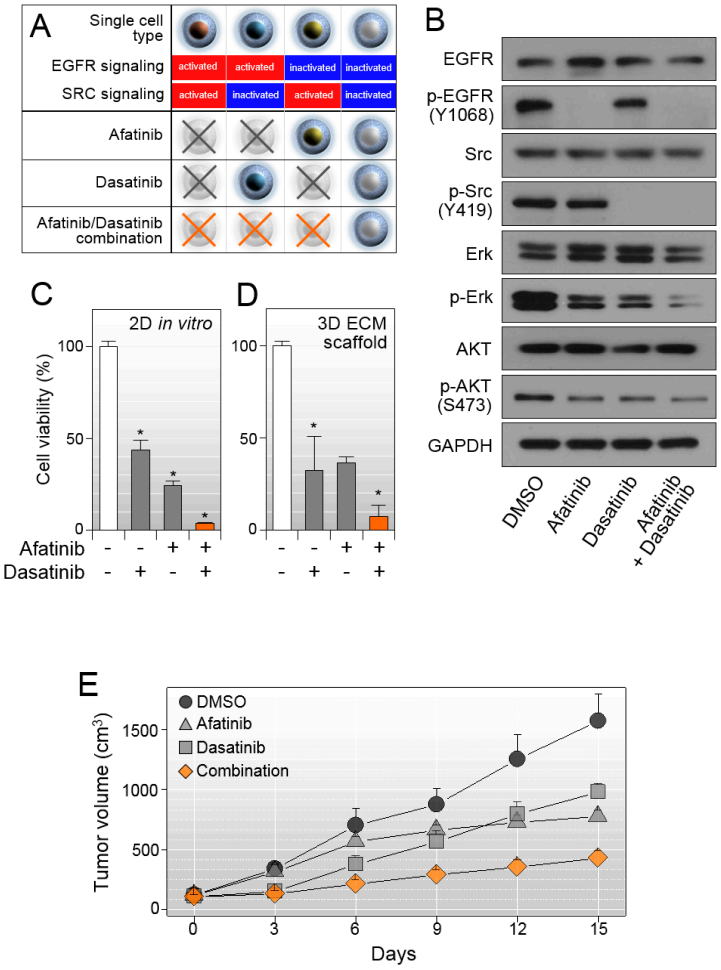


Figure 31. Combinatorial treatment of targeted drugs to check the better treatment effects for killing cancer cells. (A) Strategic assumptions of killing cancer cells based on different activation status of EGFR and SRC signaling in a cell. (B) Validation of immuno-blotting for targeted effects for Afatinib, Dasatinib and a combination of the two drugs. (A and B) GAPDH was used for a loading control. (C-D) Measurement of cellular viability upon the drug treatment in 2D in vitro (C) and 3D ECM scaffold models (D). (E) Measurement of tumor growth in mouse model along the different condition of drug treatment. Each n=5, * P<0.05.

DISCUSSION

Single-cell genome analysis enables the measurement of the extent of ITH, which may provide clues for solving problems such as cancer recurrence, metastasis, and drug resistance (85). Single-cell RNA-seq can provide integrative information for both gene expression and somatic SNVs, which makes it a comprehensive tool to connect a cell's genotype with its expression profile and phenotype. We used tumor cell-enriched PDX cells to define genomic signatures of individual tumor cells, and then verified the applicability of translating this information into biological cancer cell phenotypes such as drug responses.

When we interpret single cell RNA-seq data, the data quality needs to be considered, because of the high magnitude of amplification in the sequencing process. Sequence errors can be incorporated during the reverse transcription, cDNA amplification, and library construction processes causing false positive mutation calls. RNA editing and monoallelic expression can also cause discrepancies between SNV calls from RNA and DNA sequencing. In Part-I, we also focused on the RNA-seq SNVs that were simultaneously detected by WES and identified in more than three single cells. This approach would minimize the probability of false positive SNV calls. On the other hand, false negative SNV calls could result from missing reads at the mutant position both for DNA and RNA sequencing, which might be misinterpreted as biological heterogeneity (86). Various approaches such as Nuc-Seq which increases the starting material by using G2/M phase cells are reported to increase the genome coverage up to 91% for DNA sequencing (87). For the RNA-seq based genotype analysis, mutations in rare transcripts are most

prone to the dropout events, suggesting that RNA-seq is only suitable for the genotyping of highly expressed oncogenic driver mutations.

Despite limitations in the accuracy of single-cell RNA-seq, in this study we observed good correlations between the merged single-cell data and the bulk cell data at both the gene expression and expressed SNV levels. Once the number of single cells exceeded 30, the averaged expression levels and consensus SNVs largely recapitulated the data from bulk populations. High levels of correlation were also detected between replicate RNA-seq analyses and with the qPCR-based method. While these concordant results and overall high expression level of *KRAS* support the validity of the *KRAS* mutation calls in RNA, 12-16% of cells had insufficient read counts at the mutant position, resulting in ambiguous calls. At the single cell DNA level, we failed to determine *KRAS* mutation status in 14% of cells, which might have been influenced by normal/low copy number status (compared to 4N status in successful cases).

First, in a lung adenocarcinoma PDX model, according to the prognostic value of the activating *KRAS* mutation and RS, PDX cells with $KRAS^{G12D}$ expression and high RS would be expected to be drug resistant. Moreover, as a whole population, the PDX cells had a high variant allele frequency of $KRAS^{G12D}$ and high RS that masked the [$KRAS^{WT}$ or no KAS] and/or low RS cell types. The use of tumoricidal anti-cancer drugs with different mechanisms (cytotoxic and targeting specific signaling pathways) dramatically changed the gene expression features of the PDX cells in this study from $KRAS^{G12D}$ + high RS to $KRAS^{G12D}$ + low RS. The result was

counterintuitive, since high RS is significantly associated with worse prognosis of LUAD patients. However, in an independent PDX case, cells with a low RS also survived *in vitro* anti-cancer treatments, supporting the validity of the unexpected results.

Second, in a metastatic renal cell carcinoma PDX model, we have shown how to derive the effective therapeutics on heterogeneous cancer population by characterizing transcripome at single-cell level. Also, our study has shown comparable but enriched mutational signatures in mRCC tumors, supporting the notion that genetic variations in advanced RCC (6, 23). Integrated analysis of WES and aCGH from bulk cell population enabled to characterize evolutionary history within mRCC with driver mutations arose from pRCC. Molecular mechanisms that may affect the patient's responses to targeted agents were also analyzed by comparing the expression profiles between different tumors. With a conventional approach of sequencing bulk tumors, general features of transcriptome was identified at stochastic average of cells comprising the whole population. From dissecting transcriptome at single-cell level, however, we could achieve the better therapeutics based on the observation that drug treatment responses varies considerably across cancer cells. Considering the results of drug screening together with estimated activation status of the most targetable signaling pathways, cancer cells could be classified into 4 groups; both activated, neither activated and either activated in EGFR signaling and SRC signaling. It implied that the selected mono-therapy based on bulk profiling has a limit to diminish cells evading the relevant targeting with inactivated signaling. This knowledge enabled us to

conceive the better therapeutics of targeting both signaling pathways without chances to evade.

Taken together, this study demonstrated that gene expression and somatic SNVs of single tumor cells could be retrieved simultaneously by single-cell RNA-seq. Furthermore, the genomic data obtained could be used to elucidate potentially drug resistant subclones and to generate hypotheses on the molecular mechanisms of treatment resistance that are masked in the whole cancer cell population.

REFERENCES

1. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome biology. 2015;16(1):127. Epub 2015/06/19.
2. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153(1):17-37. Epub 2013/04/02.
3. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275-92. Epub 2011/10/18.
4. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8. Epub 1976/10/01.
5. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, et al. Inferring tumor progression from genomic heterogeneity. Genome research. 2010;20(1):68-80. Epub 2009/11/12.
6. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883-92. Epub 2012/03/09.
7. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010;467(7319):1109-13. Epub 2010/10/29.
8. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010;467(7319):1114-7. Epub 2010/10/29.
9. Zhang J, Fujimoto J, Wedge DC, Song X, Seth S, Chow CW, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014;346(6206):256-9. Epub 2014/10/11.
10. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152(4):714-26. Epub 2013/02/19.
11. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. 2013;368(13):1199-209. Epub 2013/03/15.
12. Keats JJ, Chesi M, Egan JB, Garbitt VM, Palmer SE, Braggio E, et al. Clonal competition with alternating dominance in multiple myeloma. Blood. 2012;120(5):1067-76. Epub 2012/04/14.
13. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069-75. Epub 2008/10/25.
14. The Cancer Genome Atlas Research Network. Comprehensive molecular

profiling of lung adenocarcinoma. Nature. 2014;511(7511):543-50. Epub 2014/08/01.
15. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012;150(6):1107-20. Epub 2012/09/18.
16. Youlden DR, Cramb SM, Baade PD. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. J Thorac Oncol. 2008;3(8):819-31. Epub 2008/08/02.
17. Lindeman NI, Cagle PT, Beasley MB, Chitale DA, Dacic S, Giaccone G, et al. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. J Thorac Oncol. 2013;8(7):823-59. Epub 2013/04/05.
18. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med. 2002;8(8):816-24. Epub 2002/07/16.
19. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med. 2007;356(1):11-20. Epub 2007/01/05.
20. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, et al. Three-gene prognostic classifier for early-stage non small-cell lung cancer. J Clin Oncol. 2007;25(35):5562-9. Epub 2007/12/11.
21. Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, et al. MicroRNA signature predicts survival and relapse in lung cancer. Cancer Cell. 2008;13(1):48-57. Epub 2008/01/03.
22. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64(1):9-29. Epub 2014/01/09.
23. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet. 2014;46(3):225-33. Epub 2014/02/04.
24. Fidler IJ, Kim SJ, Langley RR. The role of the organ microenvironment in the biology and therapy of cancer metastasis. Journal of cellular biochemistry. 2007;101(4):927-36. Epub 2006/12/21.
25. Ohgaki H, Kleihues P. Genetic alterations and signaling pathways in the evolution of gliomas. Cancer science. 2009;100(12):2235-41. Epub 2009/09/10.
26. Wu JM, Fackler MJ, Halushka MK, Molavi DW, Taylor ME, Teo WW, et al. Heterogeneity of breast cancer metastases: comparison of therapeutic target expression and promoter methylation between primary tumors and their multifocal metastases. Clinical cancer research : an official journal of

the American Association for Cancer Research. 2008;14(7):1938-46.
27. Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Res. 2012;72(19):4875-82.
28. Voskoglou-Nomikos T, Pater JL, Seymour L. Clinical predictive value of the in vitro cell line, human xenograft, and mouse allograft preclinical cancer models. Clinical cancer research : an official journal of the American Association for Cancer Research. 2003;9(11):4227-39.
29. Krumbach R, Schuler J, Hofmann M, Giesemann T, Fiebig HH, Beckers T. Primary resistance to cetuximab in a panel of patient-derived tumour xenograft models: activation of MET as one mechanism for drug resistance. Eur J Cancer. 2011;47(8):1231-43.
30. Tentler JJ, Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, et al. Patient-derived tumour xenografts as models for oncology drug development. Nat Rev Clin Oncol. 2012;9(6):338-50.
31. Julien S, Merino-Trigo A, Lacroix L, Pocard M, Goere D, Mariani P, et al. Characterization of a large panel of patient-derived tumor xenografts representing the clinical heterogeneity of human colorectal cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012;18(19):5314-28.
32. Hidalgo M, Bruckheimer E, Rajeshkumar NV, Garrido-Laguna I, De Oliveira E, Rubio-Viqueira B, et al. A pilot clinical study of treatment guided by personalized tumorgrafts in patients with advanced cancer. Mol Cancer Ther. 2011;10(8):1311-6.
33. Malaney P, Nicosia SV, Dave V. One mouse, one patient paradigm: New avatars of personalized cancer therapy. Cancer Lett. 2014;344(1):1-12.
34. Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. Molecular cell. 2015;58(4):598-609. Epub 2015/05/23.
35. Hong JW, Quake SR. Integrated nanoliter systems. Nature biotechnology. 2003;21(10):1179-83. Epub 2003/10/02.
36. Atencia J, Beebe DJ. Controlled microfluidic interfaces. Nature. 2005;437(7059):648-55. Epub 2005/09/30.
37. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161(5):1202-14. Epub 2015/05/23.
38. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome research. 2001;11(6):1095-9. Epub 2001/05/31.
39. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338(6114):1622-6. Epub 2012/12/22.
40. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-

length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature biotechnology. 2012;30(8):777-82. Epub 2012/07/24.
41. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome biology. 2013;14(4):R31. Epub 2013/04/19.
42. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, et al. Dynamic proteomics of individual cancer cells in response to a drug. Science. 2008;322(5907):1511-6. Epub 2008/11/22.
43. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12. Epub 2014/06/02.
44. Beasley MB, Brambilla E, Travis WD. The 2004 World Health Organization classification of lung tumors. Seminars in roentgenology. 2005;40(2):90-7. Epub 2005/05/19.
45. Joo KM, Kim J, Jin J, Kim M, Seol HJ, Muradov J, et al. Patient-specific orthotopic glioblastoma xenograft models recapitulate the histopathology and biology of human glioblastomas in situ. Cell Rep. 2013;3(1):260-73. Epub 2013/01/22.
46. Joo KM, Kim SY, Jin X, Song SY, Kong DS, Lee JI, et al. Clinical and biological implications of CD133-positive and CD133-negative cells in glioblastomas. Laboratory investigation; a journal of technical methods and pathology. 2008;88(8):808-15. Epub 2008/06/19.
47. Lee HW, Lee JI, Lee SJ, Cho HJ, Song HJ, Jeong DE, et al. Patient-Derived Xenografts from Non-Small Cell Lung Cancer Brain Metastases Are Valuable Translational Platforms for the Development of Personalized Targeted Therapy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2014.
48. Lee DW, Choi YS, Seo YJ, Lee MY, Jeon SY, Ku B, et al. High-throughput screening (HTS) of anticancer drug efficacy on a micropillar/microwell chip platform. Analytical chemistry. 2014;86(1):535-42. Epub 2013/11/10.
49. Shin Y, Han S, Jeon JS, Yamamoto K, Zervantonakis IK, Sudo R, et al. Microfluidic assay for simultaneous culture of multiple cell types on surfaces or within hydrogels. Nature protocols. 2012;7(7):1247-59. Epub 2012/06/09.
50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20.
51. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.

52. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.
53. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5(10):91. Epub 2013/10/12.
54. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9. Epub 2013/02/12.
55. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76. Epub 2012/02/04.
56. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27(19):2648-54. Epub 2011/08/11.
57. Bao L, Pu M, Messer K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics. 2014. Epub 2014/01/07.
58. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47. Epub 2015/01/22.
59. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23(6):657-63. Epub 2007/01/20.
60. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology. 2012;30(5):413-21. Epub 2012/05/01.
61. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 2012/10/30.
62. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. Epub 2011/08/06.
63. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research. 2012;22(9):1760-74. Epub 2012/09/08.
64. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93(4):641-51. Epub

2013/10/01.
65. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nature methods. 2014;11(4):396-8. Epub 2014/03/19.
66. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. Epub 2013/01/18.
67. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. 2013;4:2612. Epub 2013/10/12.
68. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(35):15449-54. Epub 2010/08/18.
69. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43-9. Epub 2013/06/25.
70. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, et al. A pathway-based classification of human breast cancer. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(15):6994-9. Epub 2010/03/26.
71. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature methods. 2014;11(1):41-6. Epub 2013/10/22.
72. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396-401. Epub 2014/06/14.
73. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic acids research. 2010;38(Database issue):D652-7.
74. Hunker CM, Galvis A, Kruk I, Giambini H, Veisaga ML, Barbieri MA. Rab5-activating protein 6, a novel endosomal protein with a role in endocytosis. Biochemical and biophysical research communications. 2006;340(3):967-75.
75. Wang L, Yamaguchi S, Burstein MD, Terashima K, Chang K, Ng HK, et al. Novel somatic and germline mutations in intracranial germ cell tumours. Nature. 2014;511(7508):241-5.
76. Sonobe M, Kobayashi M, Ishikawa M, Kikuchi R, Nakayama E, Takahashi T, et al. Impact of KRAS and EGFR gene mutations on recurrence and

survival in patients with surgically resected lung adenocarcinomas. Ann Surg Oncol. 2012;19 Suppl 3:S347-54. Epub 2011/05/25.
77. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462(7269):108-12. Epub 2009/10/23.
78. Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet. 2005;37(1):48-55. Epub 2004/12/21.
79. Huang J, Wu S, Barrera J, Matthews K, Pan D. The Hippo signaling pathway coordinately regulates cell proliferation and apoptosis by inactivating Yorkie, the Drosophila Homolog of YAP. Cell. 2005;122(3):421-34.
80. Willers H, Azzoli CG, Santivasi WL, Xia F. Basic mechanisms of therapeutic resistance to radiation and chemotherapy in lung cancer. Cancer J. 2013;19(3):200-7.
81. Engelman JA, Chen L, Tan X, Crosby K, Guimaraes AR, Upadhyay R, et al. Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. Nat Med. 2008;14(12):1351-6.
82. Janne PA, Shaw AT, Pereira JR, Jeannin G, Vansteenkiste J, Barrios C, et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. Lancet Oncol. 2013;14(1):38-47.
83. Greulich H, Chen TH, Feng W, Janne PA, Alvarez JV, Zappaterra M, et al. Oncogenic transformation by inhibitor-sensitive and -resistant EGFR mutants. PLoS Med. 2005;2(11):e313.
84. Wu JY, Wu SG, Yang CH, Gow CH, Chang YL, Yu CJ, et al. Lung cancer with epidermal growth factor receptor exon 20 mutations is associated with poor gefitinib treatment response. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008;14(15):4877-82.
85. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338-45. Epub 2013/09/21.
86. Navin NE. Cancer genomics: one cell at a time. Genome biology. 2014;15(8):452.
87. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512(7513):155-60.

국문 초록

서론: 암 환자 개인의 고유한 유전체 특성을 보다 정확히 파악하는 것은 개인 맞춤형 암 치료를 위해 필수적이다. 하지만, 일반적으로 진행되는 종양 조직을 하나의 개체로만 간주하여 유전체 특성을 파악하고 종양 내 이질성 (Intratumoral heterogeneity, ITH)를 고려하지 않는다면, 적은 분포로 존재하지만 암전이나 약물 저항성 등에 관여하는 특정 암세포 집단을 확인 하기 어렵다.

방법: 암 환자로 부터 얻은 샘플을 갖고, 종양 이질성을 극복하여 특정 암세포 집단들의 존재와 이들의 약물에 대한 반응와 기작을 확인하기 위해, 단일세포 단위로 전사체를 얻어 증폭 후 서열분석을 진행하였다.

결과: 첫째, 폐 선암 환자케이스에 적용한 경우, 각 단일세포들은 단일 염기서열 변이 (Single Nucleotide Variation, SNV)의 모자이크 패턴을 보였고, RAS-MAPK pathway 에서 다른 발현 양상을 보였다. KARS G12D 돌연변이의 유무를 확인하고, 각 세포의 폐선암에 대한 위험도를 예측하여 단일 세포들을 3 개의 특징 짓는 그룹으로 분리할 수 있었다. KRAS G12D 돌연변이를 갖고, 위험도가 높은 그룹의 경우는 RAS-MAPK pathway 가 활성화 되어 있었고, 이미 알려진 대로 Docetaxel 과 BKM120 약물에 효과가

있을 것으로 예측 되었다. 반면, KRAS G12D 돌연변이가 있음에도 위험도가 낮은 그룹의 경우는 RAS-MAPK pathway 가 상대적으로 덜 활성화 되었고, 약물에 대해서 저항성을 갖는 그룹과 비슷한 유전자 발현 양상을 보였다.

둘째, 전이성 신장암 환자 케이스에 적용한 경우, 원발암과 비교하여 전이암에서 공통적으로 나타나는 클론들이 보다 증대된 것을 관찰하였고, 전형적인 EMT 양상과 더 좋지 않은 예후의 발현 양상을 보였다. 단일세포 전사체 분석운 효과적인 약물 선택을 제시하였고, 일반적인 2 차원적 세포 배양과 종양 미세환경과 유사하게 만든 3 차원적 세포 배양 모델에서 뿐만 아니라, 쥐를 통한 동물 모델에서도 선택한 약물의 효과를 확인하였다. 나아가, 단일세포 분석을 통해 EGFR 과 Src signaling pathway 의 상호 배타적으로 활성화 된 세포의 존재를 확인하고, 이를 동시에 타겟으로 약물을 처리하여 유의한 시너지 효과를 확인하였다.

결론: 환자 유래 동물 모델을 통해 얻은 암조직의 각 단일세포들은 단일 염기서열 변이와 유전자 발현 수준에서 이질적인 성격을 보였다. 이를 통해 항암 약물에 차별적으로 반응하는 다른 그룹으로 특징 지을 수 있었다. 이러한 결과는, 단일 세포 전사체 분석을 통해 이질성의 성질로 기인하여 예측하기 힘든 종양의 특징을 보다 자세한 수준에서 파악하여, 항암치료 계획에 효과적으로 활용할 수 있음을 보여준다.

주요어 : 단일 세포 분석, 폐 선암, 신장암, 환자 유래 동물모델, 종양 이질성, 약물 반응

학 번 : 2012-30575