mitoSpliter: a mitochondrial RNA variants-based method for single-cell RNA-seq analysis of multiplexed samples

Xinrui Lin[1, *], Yingwen Chen[2, *], Li Lin[1,2, *], Kun Yin[2], Xiaoyu Wang[2], Ye Guo[2], Zhaorun Wu[2], Yingkun Zhang[2], Jin Li[3, #], Chaoyong Yang[1,2, #], Jia Song[1, #]

[1] Institute of Molecular Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China

[2] State Key Laboratory for Physical Chemistry of Solid Surfaces, Key Laboratory for Chemical Biology of Fujian Province, Key Laboratory of Analytical Chemistry, and Department of Chemical Biology, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, 361005, People's Republic of China

[3] State Key Laboratory of Genetic Engineering and School of Life Sciences, Fudan University, Shanghai, China.

* Equal contribution

[#]Correspondence: li_jin_lifescience@fudan.edu.cn (J.L.) and cyyang@xmu.edu.cn (C.Y.), songjiajia2010@shsmu.edu.cn (J.S.)

Key words: scRNA-seq; multiplexing; mitochondrial genome variants.

ABSTRACT

Single-cell RNA-seq (scRNA-seq) analysis of multiple samples separately is expensive and may cause batch effect. Multiplexing several samples for one experiment using exogenous barcodes or demultiplexing scRNA-seq data based on genome-wide RNA mutations is experimentally or computationally challenging and scale-limited, respectively. Mitochondria is a smaller genome set that can provide concise information for individual genotypes. In this study, we developed a computational algorithm "mitoSpliter" that enables sample demultiplexing based on somatic mutations in mitochondrial RNA (mtRNA). We demonstrated that mtRNA variants are computationally capable of demultiplexing large-scale scRNA-seq data. The analysis of peripheral blood mononuclear cells (PBMCs) revealed that mitoSpliter can accurately analyze up to 10 samples and 60K cells in a single experiment in ~6 hours on affordable computational resources. Further application of mitoSpliter to non-small cell lung cancer (NSCLC) cells uncovered the synthetic lethality of *TOP2A* inhibition and BET chemical degradation in the BET inhibitor-resistant cells, indicating that mitoSpliter will accelerate the application of scRNA-seq assay in biomedical research.

INTRODUCTION

In the past decades, the single-cell RNA-sequencing (scRNA-seq) assay has been broadly recognized as useful for biomedical researches[1,2]. Although the scRNA-seq assay can robustly identify single-cell clusters and cell type corresponding genetic markers[3,4], batch effects due to the variation of different rounds of experiments may introduce undesirable noises to the data set. In addition, the high cost of regular scRNA-seq experiments hesitated the application of such assays in analyzing a large number of samples.

To address these issues, several technologies have been established to enable demultiplexing the data from mixed samples with chemical probes or genetic engineering-based exogenous barcodes in the scRNA-seq experiment[5-8]. Although the data from mixed samples can be clearly and efficiently demultiplexed based on the label of chemical probes, it was still limited by the requirement of universal antibodies suitable for different analytes. The involvement of the antibodies labeling procedure also increased the cost and sample preparation time. As an alternative approach, genetic engineering-based exogenous barcodes do not require any antibodies, but their application is limited to the setting of either cell culture systems or model organisms. Thus, the acquiring of scRNA-seq data from mixed clinical samples may not be suitable to be

analyzed by these technologies in general. In addition, the experimental expertise required for creating exogenous barcodes and properly transforming them into cells prevents the global dissemination of such techniques.

Encouragingly, the demultiplexing of scRNA-seq data can also be achieved by analyzing the natural genomic variants with[9,10] or without[11-13] the data of whole genome sequencing (WGS) or whole exome sequencing (WES) as references. But these methods are computationally expensive and complicated, due to the detecting and analyzing single-nucleotide polymorphisms (SNPs) of the whole genome as endogenous barcodes. In addition, the data from either WGS or WES are still needed to assign demultiplexed cell groups to the specific samples, which may dramatically increase the cost. A demultiplexing method with comparable accuracy but significantly improved computational efficiency is desired.

Similar to the SNPs of nuclear genome, the mitochondrial genome also contains many variants. It has been validated that the variants of mitochondrial genome can be used as endogenous barcodes for lineage analysis[14-16]. We therefore hypothesized that it is possible to demultiplex the scRNA-seq data from mixed samples based on the SNPs of mitochondrial genome. In this study, we established a computational algorithm called "mitoSpliter" to demultiplex mixed scRNA-seq samples referring to the SNPs of mitochondrial RNA.

The mitoSpliter used the bulk RNA-seq data of mitochondria or whole cells, but not WGS or WES data, as references. Such a strategy dramatically reduced the experimental cost to build the reference data set and the computational cost of identifying the endogenous barcodes. With a label propagation algorithm, mitoSpliter can reliably demultiplex scRNA-seq data from either human cell lines or clinical samples with decent accuracy and computational efficiency, allowing us to analyze 60K peripheral blood mononuclear cells (PBMCs) from 10 donors in a single experiment. In addition, we applied the mitoSpliter to discover a specific synthetic lethality of Bromodomain and extra-terminal (BET) and *TOP2A* inhibition in non-small cell lung cancer (NSCLC) cell line A549.

RESULTS

Working principle of mitoSpliter

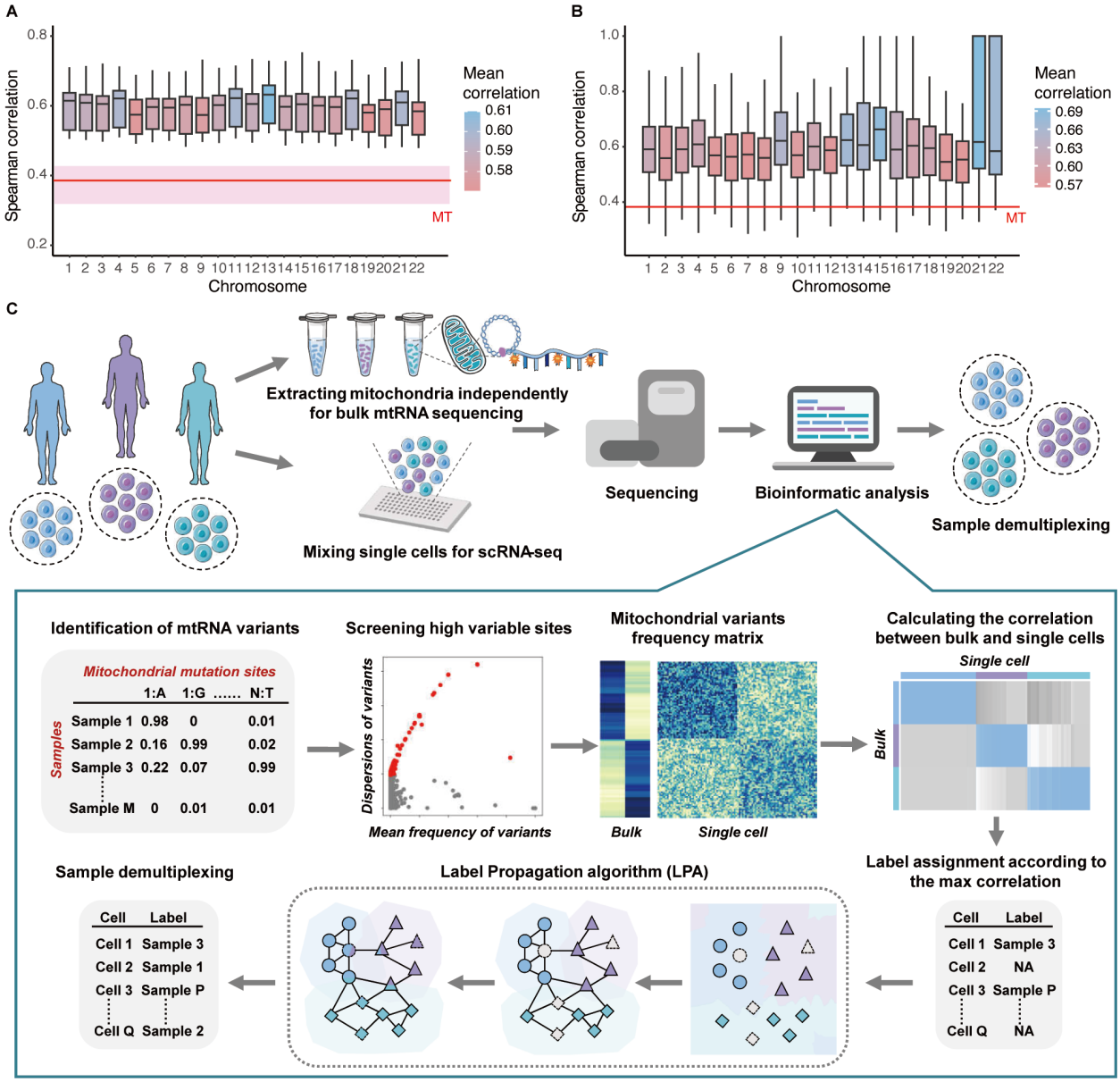

Figure 1. Mitochondrial variants are the optimal material for genotyping. (A) The Mean Spearman correlation between 5000 randomly selected pairs of two individuals in 1000 genomes project (phase 3) based on mitochondrial or autosomal variant profiles. The red line represents the mean correlation of the mitochondrial variants between individuals, the upper and lower pink boundaries represent the first and third quantiles. MT represents "mitochondrion". It revealed that mitochondrial variants profile is significantly more variable between individuals than that of other chromosomes.(B) Mean Spearman correlation between pairwise individuals from 1000 genomes project (phase 3) based on variant profiles of mitochondrial or autosomal segments. The length of an autosomal segment is equal to the length of the mitochondrial genome, and 100 segments were randomly

extracted from each autosomal. The red line represents the mean correlation of the mitochondrial genome between individuals. MT represents "mitochondrion". It revealed that mitochondrial variant profiles are significantly more variable than those of other chromosomes within the same length interval (1 mitochondrial length). (C) Overview of mitoSpliter framework. SNPs were identified by aligning RNA-seq data to the reference genome for each sample. Based on variation frequency and dispersions in bulk RNA-seq, high-variable mitochondrial SNPs were used to genotype samples. The mitochondrial genotyping correlation between bulk samples and cells was used to determine the sample label of some single cells. Only cells with high bulk sample correction can be labeled. Then, the label propagation algorithm (LPA) was employed to give sample labels to unlabeled cells based on single-cell mtRNA SNPs.

In comparison to other chromosomes, the mitochondrial genome is much shorter, with only 17K base pairs, making it feasible to achieve a greater sequencing depth than another chromosome (**Supplementary Figure 1**). The shorter length of the mitochondrial genome dramatically decreased the required computational resources, and this unique feature makes the mitochondrial genome an ideal material for genotyping. To estimate if the mitochondrial genome variants within a population are diverse enough for genotyping, we analyzed the whole genome sequencing data from the 1,000 genomes project (phase 3)[17]. We found that the average mitochondrial SNP profile variations between individuals were much greater than those of other chromosomes (**Figure 1A-B**). Therefore, we hypothesize that it is feasible to use the mtRNA variants for genotyping based on bulk RNA-seq data of either mitochondria or whole cells and it is possible to demultiplex the scRNA-seq data from mixed samples (from different donors) based on mtRNA variants.

Then, we established an algorithm called "mitoSpliter" to demonstrate these hypotheses (**Figure 1C**). Detailly, the mitochondrial SNPs were identified by aligning and analyzing the RNA-seq data (including both bulk and single-cell) to the reference genome for each sample. The high-variable mitochondrial SNPs in bulk data were then selected for mitochondrial genotyping of samples according to the variant's frequency and dispersions. Once the mitochondrial genotyping of samples was confirmed, a label representing which sample the individual cell was derived from was inferred according to the correlation among bulk samples and cells, which was then added as the metadata of the scRNA-seq dataset. Due to the low depth of single-cell sequencing, some cells are unable to receive reads mapped to the selected highly variable sites (mitochondrial genotyping sites) and therefore are unable to acquire labels. To meet this issue, a semi-supervised machine learning strategy, known as label propagation algorithm (LPA), was integrated to transfer sample labels among cells according to the mtRNA SNPs detected in single cells. Finally, the distribution

of cells from different samples was visualized by assigning different colors for the samples on the UMAP or TSNE plots. The source code and documentation for this computational algorithm, mitoSpliter, are available on Github (https://github.com/lnscan/mitoSpliter).

Validation of mitoSpliter with virtually multiplexed scRNA-seq data

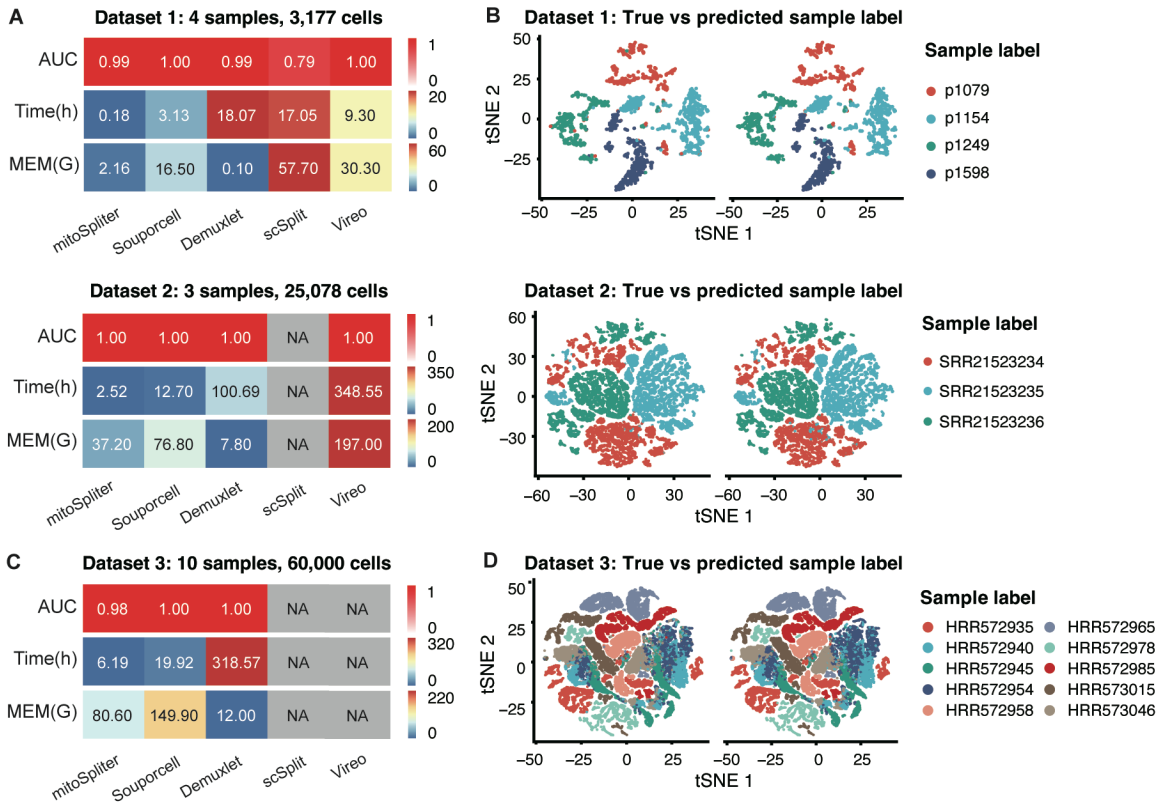

Figure 2. Performance of virtually multiplexed scRNA-seq data. (A) Comparison of mitoSpliter and other demultiplexing tools in two datasets with 4 and 3 samples. The AUC (area under the receiver operating characteristic curve), running time, and memory requirements for each tool are displayed. 'NA' indicates that the tool was unable to finish the computation with the maximal memory (600G) on our server. (B) TSNE plot based on highly variable variants identified by mitoSpliter in two datasets with 4 and 3 samples. The left subplots were colored in true sample labels, while the right subplots were colored in sample labels predicted by mitoSpliter. C) Comparison of mitoSpliter and other demultiplexing tools in a large dataset with 10 samples. The AUC, running time, and memory requirements for each tool are displayed. 'NA' indicates that the tool was unable to finish the computation with the maximal memory (600G) on our server. (D) TSNE plot based on highly variable variants identified by mitoSpliter in this large sample. The left subplot was colored in true sample labels, while the right subplot was colored in the mitoSpliter-predicted sample labels.

We first validated the robustness of mitoSpliter using two simulation data with mixing status for 4 and 3 PBMC samples (detailed in **Supplementary Table 1 and 2**). The genotypes were

derived from the pooled scRNA-seq data for each donor. The accuracy, running time, and required memory of mitoSpliter were compared to four published computational algorithms, including Demuxlet[9], Souporcell[13], scSplit[11], and vireo[12]. The mitoSpliter presented comparable or even higher accuracy (demonstrated as AUC, see the details in **Supplementary Note 1**) than other methods for both datasets. However, the computation time and memory requirement of mitoSpliter was much lower than the other methods (**Figure 2A**). The accuracy of demultiplexing by mitoSpliter was further demonstrated in mtRNA variants-based TSNE plot. The "True" label from the original datasets and the "Predicted" label from mitoSpliter showed significant agreement (**Figure 2B**).

We then tested the mitoSpliter and the other methods with a large virtually mixed scRNA-seq dataset containing 60k cells from 10 samples (**Supplementary Table 3**). MitoSpliter achieved a decent accuracy within 6 hours on a server with ~80 Gigabytes (G) of memory, making it an excellent choice for splitting large datasets. Notably, scSplit and Vireo indeed failed to analyze this large dataset with the maximal memory (600G) on our server (**Figure 2C**). The accuracy of mitoSpliter was also demonstrated in mtRNA variants based TSNE plot (**Figure 2D**). These experiments on simulation datasets indicated that mitoSpliter may be able to perform demultiplexing of label-free scRNA-seq data with minimal computational resources.

Validation of mitoSpliter with pseudo-mixed PBMC samples

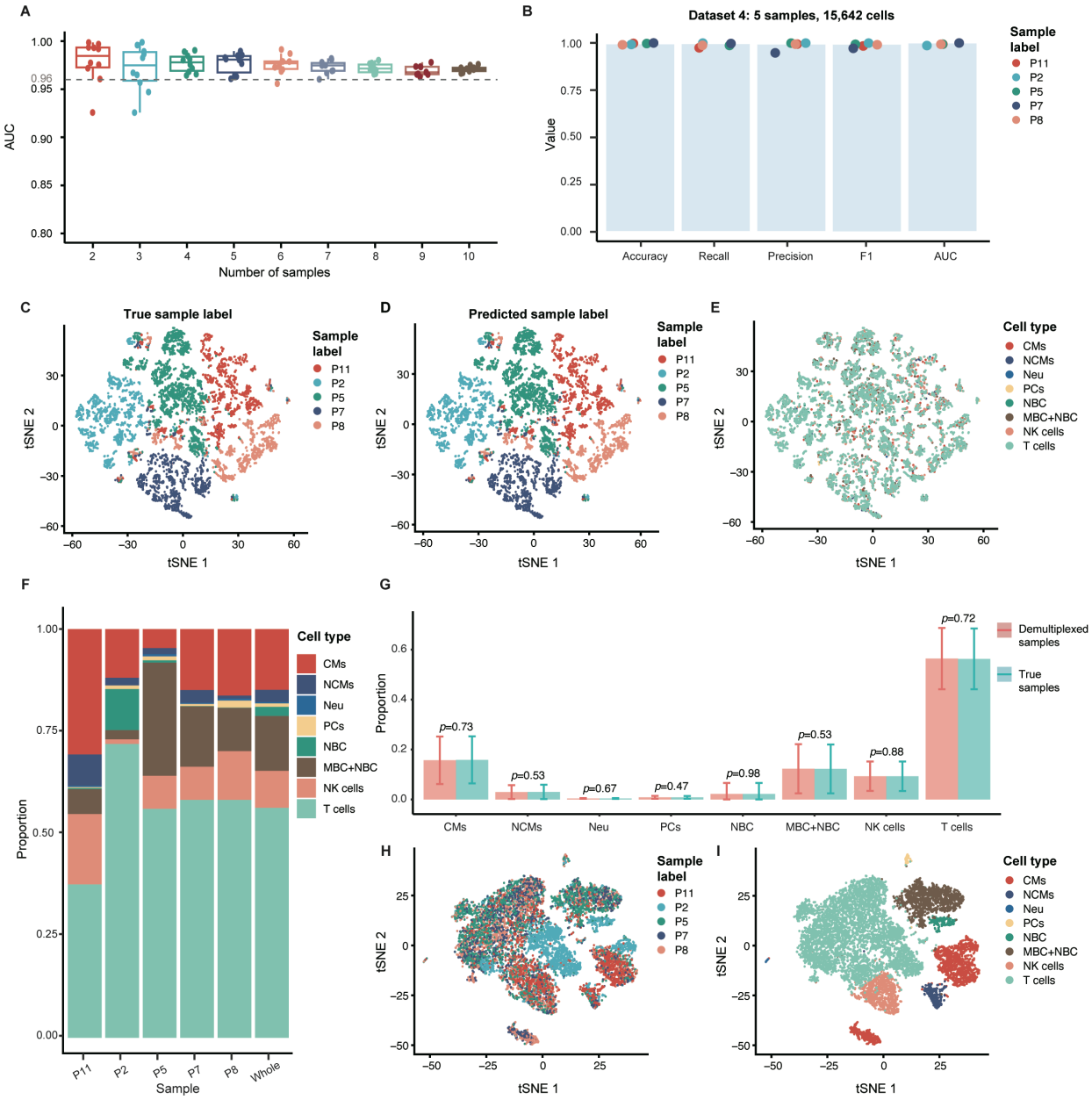

Figure 3. Performance of mitoSpliter on pseudo-mixed PBMC samples. (A) AUC values of the demultiplexing outcomes of mitoSpliter for various numbers of mixed samples. (B) A comprehensive evaluation of the mitoSpliter demultiplexing results for a dataset of five mixed samples. For each sample, the demultiplexing accuracy, recall rate, precision, *F*1 score, and AUC are presented. (C-E) TSNE plot derived from highly variable variants identified by mitoSpliter and colored based on true sample labels, predicted sample labels, and cell types, respectively. (F) The fraction of each cell type in the demultiplexed samples. (G) A comparison of the cell type composition of demultiplexed and true samples. Here, a paired t-test was employed. (H-I) A gene expression-

based TSNE plot colored in predicted sample labels and cell types, respectively. CMs, classical monocytes; NCMs, non-classical monocytes; Neu, neutrophil; PCs, plasma cells; NBC, naïve B cells; MBC, memory B cells; NK cells, natural killer cells.

We then evaluated the robustness of mitoSpliter with newly generated scRNA-seq data and paired mitochondrial bulk RNA-seq data from the PBMC sample. The mitochondrial genotypes were identified from low-depth bulk RNA-seq data of mitochondria (~3.7G per sample). The PBMC samples from 11 donors were analyzed by the Well-Paired-Seq assay[18] separately (**Supplementary Table 4**). The data from 2 to 10 donors were virtually multiplexed by randomly selecting donors. Ten datasets were created randomly for each condition. Encouragingly, the mitoSpliter achieved an average AUC ranging between 0.97 and 0.99 for various sample sizes and combinations (**Figure 3A, Supplementary Table 5**).

We then randomly selected a 5-donor pseudo-multiplexed data to demonstrate the details of the performance. Decent accuracies were derived from the analysis of mitoSpliter on all of these 5 donors (**Figure 3B**). To assess whether a similar biological conclusion can be drawn from the computationally demultiplexed data and experimentally separated data, we compared the results of the downstream analysis with the "true" label and the "predicted" label. The mtRNA variants-based TSNE plot presented the overall agreement of the data with the "true" label and the "predicted" label from mitoSpliter (**Figure 3C-D**). In addition, the universal distribution of 8 cell types from the 5 donors on the TSNE plots was observed (**Figure 3E-F**). There was no significant difference in the proportion of each cell type with the "true" label and the "predicted" label in the data from the five donors (a paired t-test was used, **Figure 3G, Supplementary Table 6**). Notably, the remarkable batch effects were observed on gene expression-based TSNE plots, which highlighted the importance of performing scRNA-seq analysis with mixed samples (**Figure 3H-I**). These results highlighted the capacity of mitoSpliter to perform genotyping based on low-depth bulk RNA-seq data of mitochondria and demultiplex scRNA-seq data from up to 10 donors.

Validation of mitoSpliter with experimentally multiplexed scRNA-seq data

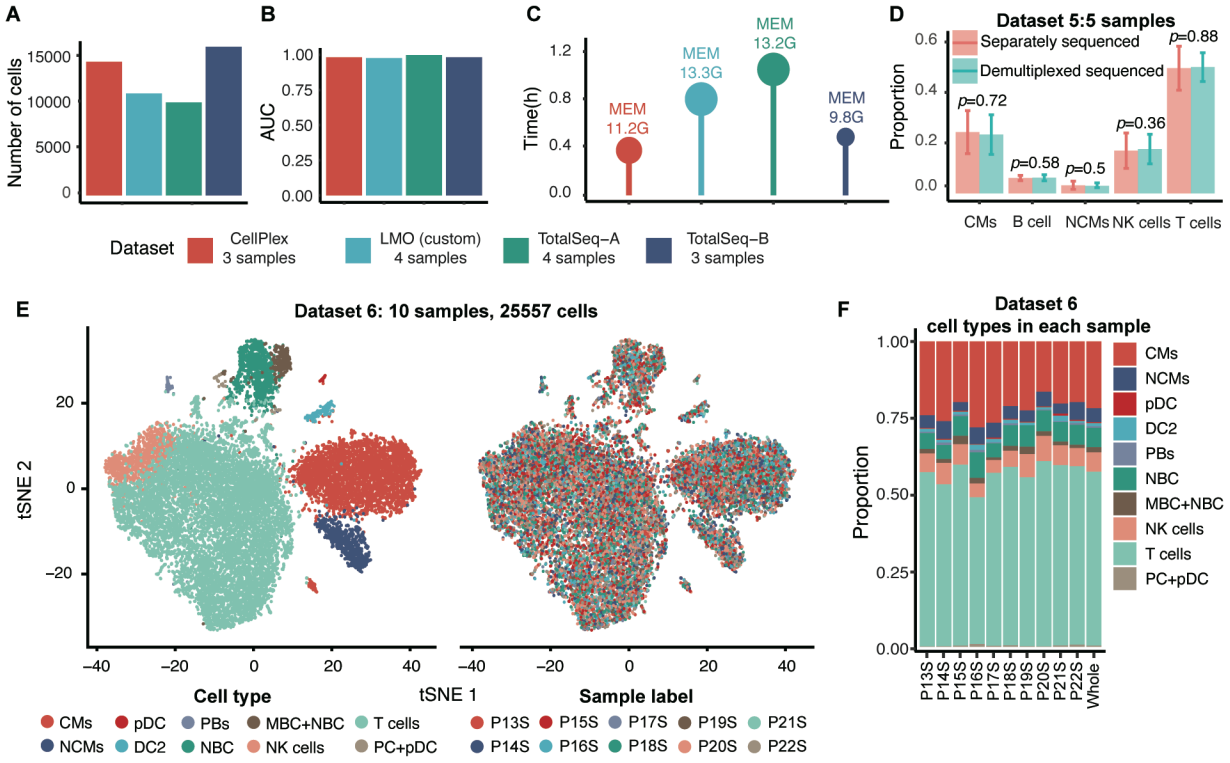

Figure 4. Performance of mitoSpliter on experimentally multiplexed scRNA-seq data. (A) Number of cells in each dataset. (B) AUC of mitoSpliter on four multiplexed datasets with exogenous barcodes. (C) Running time and memory requirement of mitoSpliter on four multiplexed datasets with exogenous barcodes. (D) A comparison of the cell type composition of demultiplexed and separately sequenced samples. Here, a paired t-test was employed. (E) Gene expression-based TSNE plot for multiplexed data with 10 samples. The left was colored in cell types, while the right was colored in predicted sample labels. (F) The proportion of each cell type in the demultiplexed samples. CMs, classical monocytes; NCMs, non-classical monocytes; pDC, plasmacytoid dendritic cells; DC2, type 2 dendritic cells; PBs, plasmablasts; NBC, naïve B cells; MBC, memory B cells; NK cells, natural killer cells; PCs, plasma cells; Mono, monocytes.

Using the exogenous barcodes as the gold standard, we then verified the demultiplexing capability of mitoSpliter with multiplexed scRNA-seq data from antibody- or lipid-based multiplexing methods (**Figure 4A, Supplementary Table 7**). Notably, mitoSpliter achieved an AUC of 98%-100% on these data (**Figure 4B**), with a maximum time and memory consumption of 1.2h and 13.3G (**Figure 4C**), demonstrating its robustness and high efficiency on experimentally mixed samples.

To further determine the robustness of mitoSpliter on primary samples, 39,106 and 31,973 PBMCs from 5 donors were analyzed in both separated and pooled fashion in parallel

(**Supplementary Table 8**). MitoSpliter completed the analysis in less than 0.37 hours with 14.4G of RAM. 5,149 ± 1,103 cells were demultiplexed per sample (**Supplementary Figure 2A**). There was also no significant difference in the proportion of each cell type among the separated sequenced and demultiplexed samples from the five donors (a paired t-test was used, **Figure 4D**). These results indicated that the demultiplexed results of mitoSpliter were comparable to those of separated analysis.

Then, we tested a large label-free scRNA-seq dataset of 25,223 PBMCs from ten donors (**Supplementary Table 9**). MitoSpliter still completed the analysis in under 1 hour while using a negligible amount of RAM (27G). 2,142 ± 922 cells were detected per sample (**Supplementary Figure 2B**). We observed minimal batch effects of different donors in the gene expression-based TSNE plot (**Figure 4E**). The proportions of each cell type in the demultiplexed data of the ten donors were consistent with the existing knowledge of PBMC (**Figure 4F**). These findings imply that mitoSpliter can reduce batch effects by combining as many as ten samples into a single scRNA-seq.

scRNA-seq analysis of NSCLC by mitoSpliter identified the synthetic lethality of BET degrader and *TOP2A* inhibitor

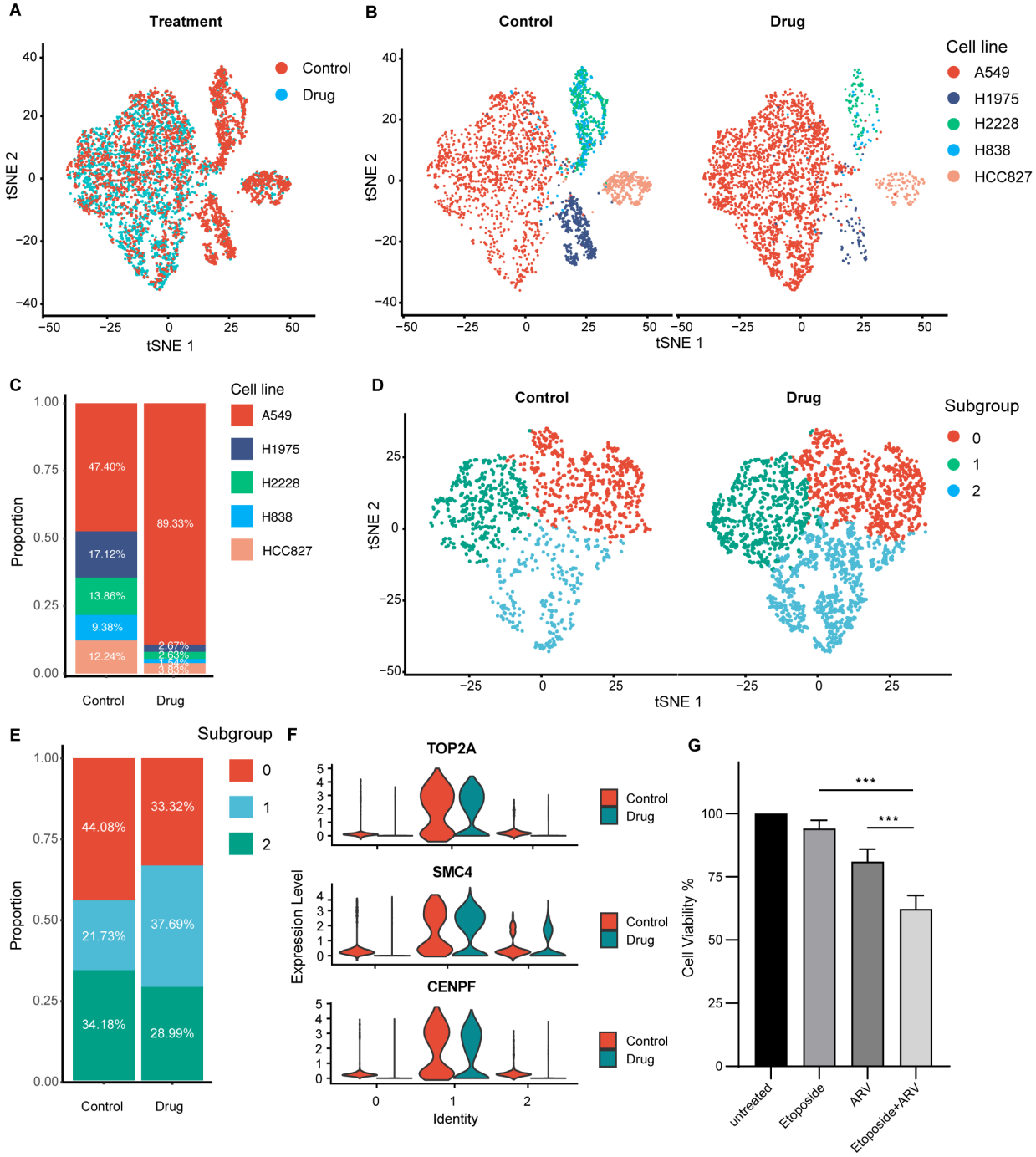

Figure 5. scRNA-seq analysis of NSCLC by mitoSpliter identified the synthetic lethality of BET degrader and *TOP2A* inhibitor. (A) A combined gene expression-based TSNE plot of NSCLC cells colored in drug treatment status. (B) Two TSNE plots of NSCLC cells based on gene expression and colored by cell lines. The cells on the left were not treated with the drug, whereas the cells on the right were. (C) The proportion of cell

lines in sequencing results before and after drug treatment. (D) Two gene expression-based TSNE plots of A549 cells. The cells on the left were not treated with the drug, whereas the cells on the right were. Cells are colored in subgroup types. (E) The proportion of subgroup cell clusters in the A549 cell line before and after drug treatment. The proportion of cells belonging to subgroup 1 increased after drug administration. (F) Expression of the top three up-regulatory genes identified in A549 subgroup 1 across all subgroups. (G) Combined administration of Etoposide and ARV can significantly reduce the A549 cell viability.

NSCLC is an important type of cancer worldwide. It has been demonstrated that the inhibition or degradation of the bromodomain and extra terminal (BET) proteins may inhibit the proliferation of NSCLC cells[19-21]. But it has also been reported multiple times that the sensitivity of NSCLC cells to the chemical inhibition or degradation of BET is highly variable[22-25]. Though several studies have been carried out to understand how the sensitivity of NSCLC cells to the BET inhibitors was regulated, the drugs which can lead to synthetic lethality for the resistant cells are still limited.

We identified that the chemical degradation of the BET via the treatment of ARV-771[26] had different effects on the proliferation rate of NSCLC cell lines (**Supplementary Figure 3A**). Interestingly, cell line A549 presented remarkable resistance to ARV-771. So, we were trying to understand the underlying mechanism of resistance with the application of mitoSpliter.

To avoid the batch effects, we performed multiplexed scRNA-seq analysis of five NSCLC cell lines, with or without the treatment of ARV-771 for 48 hours, in one experiment. Multiple clusters were identified by analysis purely based on the expression profile (**Figure 5A**). Based on the bulk RNA-seq data of mitochondria, we managed to assign the cells to their origin and present them in the TSNE plots (**Figure 5B, Supplementary Table 10**). Notably, the ARV-771 treatment decreased the number of cells except A549 cells (**Figure 5C**).

We explored the mechanism of resistance against ARV-771 by computationally isolating and reanalyzing the scRNA-seq data of A549 cells. We found the emergence of cluster 1 with ARV-771 (**Figure 5D and E**). We then identified a list of markers for these new clusters (**Supplementary Table 11**), whose functional enrichment revealed the features related to mitosis and cell cycle (**Supplementary Figure 3B and C**). Particularly, we observed the upregulation of *TOP2A* which is a famous therapeutic target for NSCLC (**Figure 5F**). Importantly, chemical inhibition of *TOP2A* via etoposide in the combination of ARV-771 synergistically prevented the proliferation of A549 cells specifically (**Figure 5G, Supplementary Figure 3D**).

DISCUSSION

In the current study, we described a method "mitoSpliter" for scRNA-seq analysis of different samples in one experiment. MitoSpliter can analyze multiple samples in one experiment based on the variants of mtRNA with reasonable running time and memory. The application of mitoSpliter on NSCLC revealed a synthetic lethality of *TOP2A* inhibition and BET chemical degradation.

The application of scRNA-seq to understand the mechanism of diseases is limited by its high cost, as the high complexities of diseases usually required the analysis of a large number of samples. One strategy is to increase the throughput by multiplexing multiple samples in one experiment[27-30]. However, the existing strategies with endogenous or exogenous barcodes required significant extra experimental or computational costs. In comparison, mitoSpliter only required low-depth sequencing data of mtRNA, which dramatically reduced the overall cost of multiplexed scRNA-seq assay. We are expecting that the mitoSpliter will be employed in projects related to epidemiology or population genetics studies.

The batch effects among different experiments are a major technical issue for scRNA-seq analysis. Despite many efforts from the community[31-35], performing multiplexed scRNA-seq analysis in one experiment is still the ideal method to solve this issue. The existing methods to demultiplex scRNA-seq data with either exogenous or endogenous barcodes require extensive experimental or computational efforts. The launch of mitoSpliter enables us to demultiplex label-free scRNA-seq data with minimal computational expertise. With mitoSpliter, biologists with limited experimental and computational capability can also analyze the mixed samples with scRNA-seq assay.

One concern we had, before we developed the mitoSpliter, is that the SNPs of the mitochondrial genome may be in lack of enough diversities to demultiplex complicated samples. Fortunately, the robustness of mitoSpliter was validated with various kinds of datasets containing 2-10 samples in our study. The mitoSpliter presented comparable or even higher accuracy in assigning cells into samples compared to other methods, though it required less RAM and running time for all the tested datasets. Notably, a similar performance of mitoSpliter was observed on scRNA-seq data from either cell lines or primary cells.

The BET inhibitor or chemical degrader has been recognized as a potential anti-cancer reagent for a broad range of cancer cells. Thus, it is important to understand the mechanism for the resistance to ARV-771 in A549 cells. Our finding revealed a clonal evolution event of A549 cells

upon the treatment of ARV-771, which brought us to identify a previously unknown synthetic lethality of BET degrader and *TOP2A* inhibitor. The detailed mechanism for the emergence of the resistant clone still needs further studies.

The mitoSpliter is certainly limited by the requirement of mitochondrial genome information in the scRNA-seq data. Therefore, it is not suitable to apply mitoSpliter on single nucleus RNA-seq (snRNA-seq) analysis. Since the snRNA-seq is becoming a popular assay to analyze frozen clinical samples, methods specially designed for demultiplexing snRNA-seq data are going to be established in the future.

In conclusion, we created the mitoSpliter assay for analyzing multiplexed scRNA-seq data reliably and efficiently. With mitoSpliter, we removed most of the technical hurdles for the biologists to perform scRNA-seq experiments with mixed samples. The mitoSpliter will further accelerate the application of scRNA-seq assay in biomedical research.

300 **MATERIAL AND METHODS**

301 **Cell preparation**

302 The whole blood samples were obtained from donors, which was approved by the ethical
303 committee of Xiamen University (Protocol No. KYX-2018-006). Briefly, 3 mL whole blood was
304 fully mixed with 3 mL 1×PBS buffer and subsequently added to the upper of 6 mL Ficoll-Paque
305 Plus (GE Healthcare, 17-1440-02) carefully. After centrifugation for 30 min at 2000 rpm at 15°C,
306 the PBMC cell layer was carefully aspirated, transferred to a new tube, and washed twice with
307 1×PBS buffer. The supernatant was removed and 300μL RPMI-1640 (Hyclone, SH30809.01) with
308 10% FBS (Viva cell, C04001-500) and 900μL red blood cell lysis buffer (Solarbio, R1010) were
309 added to the pellet, then the sample was placed on ice for 15min. After centrifugation for 3min at
310 2000 rpm at 4°C and supernatant removal, cells were washed once with 1×PBS buffer and
311 resuspended in 1×PBS buffer for the mitoSpliter experiment. The cell viability was confirmed by
312 Countess™ Cell Counting Chamber Slides (Invitrogen, C10283).

313 H2228, HCC827, H1975, A549, and H838 cells were cultured in RPMI-1640 supplemented
314 with 10% FBS and 1% penicillin-streptomycin (Invitrogen, 15140122) at 37°C and 5% CO_2. For
315 preparing scRNA-seq samples for multiplexed analysis, the five cell lines were co-cultured with
316 equal cell input in two culture plates, one was subject to 0.2 μM ARV-771 (MedChemExpress,
317 HY-100972) treatment for 48 hours and another was cultured without drug treatment. For
318 preparing bulk mtRNA samples, cells from the five cell lines were cultured independently without
319 drug treatment. For cell collection, the media was removed, then cells were rinsed with 1×PBS
320 buffer and detached from the surface of the cultured plates by applying Trypsin (Gibco, 25200072)
321 and incubated at room temperature for 2min. After cells were deadhered, cells were harvested by
322 rising with complete media and washed three times with 1×PBS buffer, then cell pellets were
323 resuspended in 1×PBS buffer for the mitoSpliter experiment.

324 **Sequencing library preparation of mitoSpliter**

325 For bulk mitochondrial RNA (mtRNA) sequencing, the Cell Mitochondria Isolation Kit
326 (Beyotime, C3601) was first used to isolate the mitochondria of individual samples, the operation
327 followed manual instruction. Then Total DNA/RNA Isolation Kit (OMEGA, R6731-01) was used
328 to extract and purify RNA from the isolated mitochondria. The library preparation for bulk mtRNA
329 sequencing followed a modified Smart-Seq2 protocol[36,37]. The purified mitochondrial RNA (5 μL)
330 was subject to reverse transcription by adding 2 μL 10 μM oligo-dT (ordered from Sangon), 2 μL

331 10 mM dNTPs (TransGen Biotech, AD101-11), 4 μL 5× RT buffer, 1 μL Maxima H Minus reverse
332 transcriptase (Thermo Scientific, EP0752), 1 μL 50 μM Template Switch oligo (ordered from
333 Sangon), 1 μL RNase Inhibitor (TransGen Biotech, AI101-02), 2 μL 10 mM GTP (Invitrogen,
334 R0461), and 2 μL 50% PEG-8000 (Beyotime, R0056). The reverse transcription was incubated at
335 42°C for 90 min followed by 5 cycles of 50°C for 2 min and 42°C for 2 min, and a final incubation
336 at 70°C for 15 min. For PCR, 25 μL 2×HiFi HotStart Readymix (Kapa Biosystems, KK2602), 0.8
337 μL 50 μM ISPCR oligo (ordered from Sangon), and 4.2 μL Nuclease-free water were added. The
338 PCR reaction was incubated at 98°C for 3min followed by 10 cycles of 98°C for 20 s, 67°C for 15
339 s, and 72°C for 6 min, and a final incubation at 72°C for 6min. Cells from different donors or cell
340 lines were independently subject to the procedure above.

341 For single-cell RNA sequencing, the chip fabrication, barcode beads preparation, and
342 sequencing library preparation were performed as described previously[18]. To prepare pseudo-
343 mixed PBMC samples, cells from 11 donors were subject to scRNA-seq independently, and 10000-
344 dual-well chips were used. To prepare experimentally multiplexed scRNA-seq data, PBMCs from
345 5 donors were subject to independent and multiplexed scRNA-seq (cells from different donors
346 were mixed thoroughly with the equal input before loading to the Well-paired-seq chip)
347 simultaneously, and 30000-dual-well chips were used. To generate the large scRNA-seq dataset
348 from 10 donors, PBMCs from 10 donors were subject to multiplexed scRNA-seq, and 100000-
349 dual-well chips were used for scRNA-seq.

350 The amplification products of both bulk and scRNA-seq samples were purified twice with 0.6×
351 VAHTS DNA Clean Beads (Vazyme, N411-02), and cDNA concentrations were measured with a
352 Qubit fluorometer. An appropriate volume of cDNA was input for library construction using
353 TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme, cat#TD502). The libraries were purified
354 twice with 0.6× VAHTS DNA Clean Beads and cDNA concentration was measured again. The
355 libraries were sequenced on Illumina Novaseq 6000 instruments with paired-end 150 base reads.

356 **Identification of different cell proliferation rates of NSCLC cell lines under the treatment of**
357 **ARV-771**

358 The cultured H2228, HCC827, H1975, A549 and H838 cells were harvested independently
359 and diluted to 5 cells/μL with complete culture media, then 100 μL cell suspension of each cell
360 line were added to each well of different 96-well plates. After a 24-hour culture for cell adhesion,

361 we exposed each cell line to ARV-771 at six doses (0.0625 μM, 0.125 μM, 0.25 μM, 0.5 μM, 1
362 μM, and 2 μM) in six duplicates. These conditions and vehicle control accounted for 42
363 independently treated cell populations of each cell line in this experiment. After a 48-hour
364 treatment, cell counting kit-8 (Apexbio, K1018-5) assay was used to measure the cell viability of
365 each condition, and the values of IC_{50} of each cell line were calculated to confirm the best ARV-
366 771 concentration where cells of NSCLC cell lines would present different cell proliferation rates.

367 **Validation of the effect of the combined use of ARV-771 and etoposide on the proliferation of**
368 **A549 cells**

369 Pre-experiments were performed to confirm a dose range that etoposide has inhibited effect on
370 the proliferation rate of A549 cells, then we exposed A549 cells to ARV-771 and etoposide
371 (MedChemExpress, HY-13629) at the following combination of dose, 0 μM, 0.0625 μM, 0.125
372 μM, 0.25 μM, 0.5 μM, 1 μM and 2 μM of ARV-771, and 0 μM, 0.01 μM, 0.1 μM, 0.2 μM, 0.4 μM,
373 0.8 μM, 1 μM of etoposide. These combinations and vehicle control accounted for 48
374 independently treated cell populations in this experiment, and each combination had six replicates.
375 CCK-8 was used to measure the cell viability of each condition, and the combination of 0.5 μM
376 ARV-771 and 0.2 μM etoposide was chosen to perform statistical analysis with controls.

377 **Data pre-processing.**

378 To verify the reliability and performance of mitoSpliter, we firstly collected several publicly
379 available single-cell sequencing data.

380 General single-cell sequencing data included dataset 1, which consisted of 3,177 PBMCs from
381 4 individuals (detailed in **Supplementary Table 1**) obtained from Gene Expression Omnibus
382 (GEO) with accession number GSE96583[9], dataset 2, which consisted of 3 individuals of 25,078
383 PBMCs (detailed in **Supplementary Table 2**) obtained from GSE213096, and dataset 3, which
384 consisted of 10 individuals with 60,000 PBMCs (detailed in **Supplementary Table 3**) obtained
385 from the GSA-Human repository with accession number HRA001748.

386 scRNA-seq data with exogenous barcodes were obtained from a previous study[38] (detailed in
387 **Supplementary Table 7**). Datasets generated by CellPlex (European Nucleotide Archive (ENA)
388 repository with accession number E-MTAB-9964), covering 3 samples of 14,606 cells, LMO
389 (custom) (ENA, E-MTAB-11401), covering 4 cell lines of 11,156 cells, TotalSeq-A (ENA, E-
390 MTAB-11401), covering 4 cell lines of 10,171 cells, and TotalSeq-B (ENA, E-MTAB-9964),
391 covering 3 samples of 16,263 cells were downloaded.

For four sets of sequencing data, mitochondrial bulk RNA-seq were generated using a modified Smart-Seq2 protocol[36,37] and single-cell RNA-seq were generated from the Well-Paired-Seq[18] platform. The first set covered 42,779 PBMCs from 11 individuals (dataset 4, detailed in **Supplementary Table 4**); the second set of PBMCs from 5 donors contained independent mitochondrial bulk RNA-seq and single-cell RNA-seq data (included in dataset 5, detailed in **Supplementary Table 8**), as well as a multiplexed single-cell RNA-seq data with cells from these 5 donors (included in dataset 5, detailed in **Supplementary Table 8**); the third set is a multiplexed single-cell RNA-seq data covering 25,557 PBMCs from 10 donors (dataset 6, detailed in **Supplementary Table 9**); the fourth set consisted of two multiplexed single-cell RNA-seq data covering 5 non-small cell lung cancer cell lines under drug and control treatment, with 3,229 drug-treated cells and 3,126 control cells sequenced (detailed in **Supplementary Table 10**).

Generation of single-cell and bulk mitochondrial mutation matrix

Then, three steps were required to generate single-cell and bulk mitochondrial mutation matrices: sequence alignment, remapping, and variant calling. In the process of sequence alignment process, scRNA-seq data were mapped to GRCh38 genome with CellRanger v3.1.0 (for 10x Genomic RNA-seq data) or zUMIs v2.9.7e[39] (for Well-Paired-Seq data), and bulk mitochondria RNA-seq data were quality-filtered using cutadapt v1.18[40] and aligned to the mitochondrial sequence in GRCh38 genome using STAR v2.7.3a[41]. Remapping of obtained bam files cound improve the quality of the alignment and reduce the false positive rate of variant calling. After the bam file was remapped to the reference genome using minimap2 v2.24[42] （parameters: -ax splice -t 8 -G50k -k 21 -w 11 --sr -A2 -B8 -O12,32 -E2,1 -r200 -p.5 -N20 -f1000,5000 -n2 -m20 -s40 -g2000 -2K50m --secondary=no）, variant calling was performed using a python mtDNA genotyping pipeline[16]. Detailly, reads in the remapping single-cell bam file with base quality or alignment quality lower than 20 were filtered, and then the allele frequency of each cell was extracted. For the bulk mitochondria RNA-seq bam file, we set the threshold value of base quality and alignment quality to 10. Finally, we integrated all of the mutations from all samples or cells to create a bulk or single-cell mutation matrix.

Demultiplexing of mixed sequencing data with mitoSpliter

Given a bulk mitochondrial mutation matrix, the high variable SNPs across bulk samples were selected using scanpy v1.9.1[43] for mitochondrial genotyping. As these mitochondrial sites were

more donor-specific, we calculated the correlation between single cells and each bulk sample based on the variants frequency of these sites to evaluate the probability of cells from each donor. A single cell was considered to be from the same donor as the bulk sample whose bulk sample exhibited the highest correlation unless it lacks a positively correlated bulk sample. Due to the uneven coverage of scRNA-seq data, cells that lacked mitochondrial genotyping sites were unable to receive donor labels through the aforementioned pipeline. Based on the mtRNA SNPs in scRNA-seq and the inferred labels, we then predicted the donor origin of each unlabeled cell. We defined this task as a semi-supervised learning problem and integrated LPA to propagate labels through cells using mtRNA SNPs in scRNA-seq.

Detailly, depending on whether the sample source labels were obtained, we classified (X, Y) into labeled data as $(x_1, y_1) \ldots (x_l, y_l)$ and unlabeled data as $(x_{l+1}, y_{l+1}) \ldots (x_{l+u}, y_{l+u})$, where X was the matrix of mtRNA SNPs, Y was the matrix of cell sample source label, $\mathbf{Y_L}=\{y_1, \ldots y_l\}$ consisted of inferred labels and $\mathbf{Y_U}=\{y_{l+1} \ldots y_{l+u}\}$ was undetermined. Thus, the task was to estimate $\mathbf{Y_U}$ from $\mathbf{X}$ and $\mathbf{Y_L}$. LPA first created a fully connected graph according to the similarity matrix constructed from $\mathbf{X}$, where the nodes represented all cells. The edges between nodes were weighted by calculating the normalized graph Laplacian matrix. Then labeled nodes acted as sources to spread labels $\mathbf{Y_L}$ among cells through the edges. Larger edge weights meant easier label propagation. Finally, unlabeled cells got persistent sample labels $\mathbf{Y_U}$ when the network of graph reached a globally stable state. In addition, we set clamping factor α=0.8, which helped the algorithm kept 80% of input labels and changed the confidence of labels within 20% during propagation. It's an optimization of inferred labels based on unlabeled data. Based on the LPA model, the probability of each unlabeled cell belonging to each donor would be calculated. The final cell sample label is determined by the LPA result with a probability greater than $(1\text{-}10^{-10})$.

Data dimension reduction and visualization.

With mtRNA mutation-based tSNE plots, the effect of mitochondrial mutation on demultiplexing was illustrated. Similar to the workflow for single-cell gene expression, Seurat was used for data normalization, scaling, clustering, and dimension reduction based on the highly variable mtRNA SNP features.

Generation of simulated multiplexed single-cell and bulk RNA-seq data

To simulate multiplexed single-cell RNA-seq data, we combined multiple single-cell

sequencing data. Based on the number of mixed samples *k* and the number of singlets predicted by scrublet v0.2.3[49] in the mixed samples, we first randomly selected the specified number of cells within each sample, followed by the random selection of 1/*k* of the reads from their corresponding reads for mixing. Reads with the same read ID from different samples were removed using seqkit v2.3.0[44] rmdup function. Then, the remaining reads were mapped to GRCh38 genome using CellRanger v3.1.0 or zUMIs v2.9.7e[39]. Single-cell mtRNA mutation matrix could be obtained after remapping and variant calling. The bulk mtRNA mutation matrix was produced by combining the reads of all cells within a sample.

Computational benchmarking.

We compared 4 demultiplexing algorithms with mitoSpliter, which were Souporcell v2.0[13], demuxlet v2[9], scSplit v1.0.9[11], and Vireo v0.2.3[12]. The maximum number of threads utilized by all algorithms was 50. Default pipeline and parameters were used for all demultiplexing calculations. For Souporcell, souporcell_pipeline.py were run through the singularity container. For demuxlet, single-cell RNA-seq data from each sample were mapped to GRCh38 genome using CellRanger v3.1.0 and bam files used for genotype calling were generated. The genotypes were then output through VCF files by samtools v1.9[45] and bcftools v1.8[45], and VCFtools v0.1.17[46] was used to merge the files into a single file. The demultiplexing results were obtained using popscle v0.1-beta demuxlet function, with merged VCF file and "GT" field. For scSplit, quality-filtering was conducted using samtools with "-q 10" and umi_tools v1.1.2[47] dedup function. Variant calling was performed using freebayes v1.2.7[48] with "-iXu -C 2 -q 1" and bcftools filter with "-I '%QUAL>30'". scSplit count matrix was generated using scSplit count function and demultiplexing was conducted using scSplit run function with "-d 0".

To conducted performance verification on pooled single-cell sequencing data, AUC value, accuracy, and recall of single cells were calculated using python package sklearn v1.0.2. Singlets predicted by scrublet v0.2.3[49] in each sample were used for performance evaluation. The running time and memory requirements of the algorithms were recorded by custom scripts.

Cell clustering and annotation

Seurat v4.3.0[50] was used to filter cells with a high proportion of mitochondrial genes or a low nFeature_RNA, and to perform data normalization, feature selection, scaling, clustering and dimension reduction. Then, cell type annotation was conducted using celltypist v1.3.0[51] with the "Immune_All_Low.pkl" model. Each Seurat cluster was annotated with the top one or two cell

types by proportion. Marker genes for each cell type were identified using "FindConservedMarkers" function in Seurat. For each cell type, paired t-test was used to compare cell type proportions across samples based on the true and predicted labels.

ACKNOWLEDGEMENTS

This work was supported by MOST 2020YFA0803601, 2018YFA0801300, NSFC 32071138 and SKLGE-2118 to Jin Li, NSFC 22104080 to Jia Song, and NSFC 21735004, 21927806 to Chaoyong Yang, and Innovative research team of high-level local universities in Shanghai SHSMU-ZLCX20212601.

AUTHOR CONTRIBUTIONS

Conceptualization, J.S., C.Y., J.L.; Investigation, L.L., Y.C.; Writing, J.S., X.L., Y.C., L.L., J.L.; Supervision, J.S., J.L., C.Y., X.L., Y.C., L.L., K.Y., X.W., Y.G., Z.W., Y.Z.

CONFLICT OF INTEREST STATEMENTS

The authors claim no conflict of interests.

REFERENCE

1 Li, J. *et al.* Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep* **17**, 178-187, doi:10.15252/embr.201540946 (2016).

2 Li, J. *et al.* Neurotensin is an anti-thermogenic peptide produced by lymphatic endothelial cells. *Cell Metab* **33**, 1449-1465 e1446, doi:10.1016/j.cmet.2021.04.019 (2021).

3 Song, J. *et al.* Entropy subspace separation-based clustering for noise reduction (ENCORE) of scRNA-seq data. *Nucleic Acids Res* **49**, e18, doi:10.1093/nar/gkaa1157 (2021).

4 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

5 Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* **19**, 224, doi:10.1186/s13059-018-1603-1 (2018).

6 McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **16**, 619-626, doi:10.1038/s41592-019-0433-8 (2019).

7 Shin, D., Lee, W., Lee, J. H. & Bang, D. Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. *Sci Adv* **5**, eaav2249, doi:10.1126/sciadv.aav2249 (2019).

8 Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods* **18**, 635-642, doi:10.1038/s41592-021-01153-z (2021).

9 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).

10 McFarland, J. M. *et al.* Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat Commun* **11**, 4296, doi:10.1038/s41467-020-17440-w (2020).

11 Xu, J. *et al.* Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol* **20**, 290, doi:10.1186/s13059-019-1852-7 (2019).
12 Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol* **20**, 273, doi:10.1186/s13059-019-1865-2 (2019).
13 Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat Methods* **17**, 615-620, doi:10.1038/s41592-020-0820-1 (2020).
14 Lin, L. *et al.* Single-cell transcriptome lineage tracing of human pancreatic development identifies distinct developmental trajectories of alpha and beta cells. *BioRxiv*, doi:doi: https://doi.org/10.1101/2021.01.14.426320 (2021).
15 Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat Biotechnol*, doi:10.1038/s41587-020-0645-6 (2020).
16 Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339 e1322, doi:10.1016/j.cell.2019.01.022 (2019).
17 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
18 Yin, K. *et al.* Well-Paired-Seq: A Size-Exclusion and Locally Quasi-Static Hydrodynamic Microwell Chip for Single-Cell RNA-Seq. *Small Methods* **6**, e2200341, doi:10.1002/smtd.202200341 (2022).
19 Gao, Z. *et al.* Targeting BRD4 proteins suppresses the growth of NSCLC through downregulation of eIF4E expression. *Cancer Biol Ther* **19**, 407-415, doi:10.1080/15384047.2018.1423923 (2018).
20 Takashima, Y. *et al.* Bromodomain and extraterminal domain inhibition synergizes with WEE1-inhibitor AZD1775 effect by impairing nonhomologous end joining and enhancing DNA damage in nonsmall cell lung cancer. *Int J Cancer* **146**, 1114-1124, doi:10.1002/ijc.32515 (2020).
21 Liu, C. *et al.* The novel BET degrader, QCA570, is highly active against the growth of human NSCLC cells and synergizes with osimertinib in suppressing osimertinib-resistant EGFR-mutant NSCLC cells. *Am J Cancer Res* **12**, 779-792 (2022).
22 Shimamura, T. *et al.* Efficacy of BET bromodomain inhibition in Kras-mutant non-small cell lung cancer. *Clin Cancer Res* **19**, 6183-6192, doi:10.1158/1078-0432.CCR-12-3904 (2013).
23 Lockwood, W. W., Zejnullahu, K., Bradner, J. E. & Varmus, H. Sensitivity of human lung adenocarcinoma cell lines to targeted inhibition of BET epigenetic signaling proteins. *Proc Natl Acad Sci U S A* **109**, 19408-19413, doi:10.1073/pnas.1216363109 (2012).
24 Gobbi, G. *et al.* The Hippo pathway modulates resistance to BET proteins inhibitors in lung cancer cells. *Oncogene* **38**, 6801-6817, doi:10.1038/s41388-019-0924-1 (2019).
25 Guo, J. *et al.* BCL6 confers KRAS-mutant non-small-cell lung cancer resistance to BET inhibitors. *J Clin Invest* **131**, doi:10.1172/JCI133090 (2021).
26 Raina, K. *et al.* PROTAC-induced BET protein degradation as a therapy for castration-resistant prostate cancer. *Proc Natl Acad Sci U S A* **113**, 7124-7129, doi:10.1073/pnas.1521738113 (2016).
27 Gupta, A. *et al.* Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Res* **32**, 242-257, doi:10.1101/gr.275509.121 (2022).
28 Yoon, Y. J. *et al.* Salivary gland organoid culture maintains distinct glandular properties of murine and human major salivary glands. *Nat Commun* **13**, 3291, doi:10.1038/s41467-022-30934-z (2022).
29 Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290, doi:10.1126/science.abl4290 (2022).
30 julian.knight@well.ox.ac.uk, C. O.-M.-o. B. A. C. E. a. & Consortium, C. O.-M.-o. B. A. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916-938 e958, doi:10.1016/j.cell.2022.01.012 (2022).

31 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420, doi:10.1038/nbt.4096 (2018).
32 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296, doi:10.1186/s13059-019-1874-1 (2019).
33 Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**, 584-586, doi:10.1038/nmeth.4263 (2017).
34 Grun, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat Methods* **11**, 637-640, doi:10.1038/nmeth.2930 (2014).
35 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058, doi:10.1038/s41592-018-0229-2 (2018).
36 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).
37 Picelli, S. Full-Length Single-Cell RNA Sequencing with Smart-seq2. *Methods Mol Biol* **1979**, 25-44, doi:10.1007/978-1-4939-9240-9_3 (2019).
38 Mylka, V. *et al.* Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq. *Genome Biol* **23**, 55, doi:10.1186/s13059-022-02628-8 (2022).
39 Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, doi:10.1093/gigascience/giy059 (2018).
40 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet. Journal* **17**, 10-12 (2011).
41 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
42 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
43 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).
44 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962, doi:10.1371/journal.pone.0163962 (2016).
45 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).
46 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).
47 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).
48 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *Quantitative Biology* (2012).
49 Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291 e289, doi:10.1016/j.cels.2018.11.005 (2019).
50 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).
51 Dominguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197, doi:10.1126/science.abl5197 (2022).