mitoSplitter: A mitochondrial variants-based method for efficient demultiplexing of pooled single-cell RNA-seq

Xinrui Lin[1, *], Yingwen Chen[2, *], Li Lin[1,2, *], Kun Yin[2], Rui Cheng[3], Xiaoyu Wang[2], Ye Guo[2], Zhaorun Wu[2], Yingkun Zhang[2], Jin Li[4, #], 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] School of Life Sciences, Shanghai University, Shanghai, 200444, China

[4] 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 can be costly and lead to batch effects. Exogenous barcodes or genome-wide RNA mutations can be used to demultiplex pooled scRNA-seq data, but they are experimentally or computationally challenging and limited in scope. Mitochondrial genomes are small but diverse, providing concise genotype information. In this study, we developed a computational algorithm called "mitoSplitter" that enables sample demultiplexing based on somatic mutations in mitochondrial RNA (mtRNA) and demonstrated that mtRNA variants can be computationally used to demultiplex large-scale scRNA-seq data. Using affordable computational resources, mitoSplitter can accurately analyze up to 10 samples and 60,000 cells in a single experiment in about 6 hours. To avoid the batch effects derived from separated experiments, we applied mitoSplitter to analyze the responses of five non-small cell lung cancer (NSCLC) cell lines to BET chemical degradation in a multiplexed fashion. We found the synthetic lethality of *TOP2A* inhibition and BET chemical degradation in BET inhibitor-resistant cells. The result indicates that mitoSplitter can accelerate the application of scRNA-seq assays in biomedical research.

Introduction

In recent decades, single-cell RNA-sequencing (scRNA-seq) assays have become widely recognized as useful for biomedical research[1,2]. While scRNA-seq assays can robustly identify single-cell clusters and genetic markers corresponding to cell types[3,4], the variation in different rounds of experiments can introduce undesirable noise to the data set, also known as batch effects. Moreover, the high cost of regular scRNA-seq experiments has limited their application in analyzing a large number of samples.

To address these challenges, various technologies have been developed to demultiplex data from mixed samples in scRNA-seq experiments using chemical probes or genetic engineering-based exogenous barcodes[5-8]. However, while the use of chemical probes can effectively demultiplex data based on labelling, it is limited by the need for universal antibodies suitable for various analytes, which also increases the cost and sample preparation time. On the other hand, genetic engineering-based exogenous barcodes do not require any antibodies, but their application is limited to cell culture systems or model organisms, making them unsuitable for acquiring scRNA-seq data from mixed clinical samples. Furthermore, the experimental expertise required

for creating exogenous barcodes and properly transforming them into cells hinders the widespread adoption of these techniques.

It is encouraging to note that the demultiplexing of scRNA-seq data can also be achieved by analyzing the natural genomic variants with[9,10] or without[11-13] the use of whole genome sequencing (WGS) or whole exome sequencing (WES) as references. However, these methods are computationally expensive and complex due to the detection and analysis of single-nucleotide polymorphisms (SNPs) of the whole genome as endogenous barcodes. Additionally, assigning demultiplexed cell groups to specific samples still requires data from either WGS or WES, which can significantly increase costs. Therefore, there is a need for a demultiplexing method that offers comparable accuracy but with significantly improved computational efficiency.

Similar to the SNPs found in the nuclear genome, the mitochondrial genome also contains many variants. It has been validated recently by our group that mitochondrial genome variants can be used as endogenous barcodes for lineage analysis[14-16]. Based on this, we hypothesized that it would be feasible to demultiplex scRNA-seq data from mixed samples using the SNPs found in the mitochondrial genome. In this study, we developed a computational algorithm called "mitoSplitter" to demultiplex mixed scRNA-seq samples by referencing the SNPs found in mitochondrial RNA.

The mitoSplitter algorithm uses bulk RNA-seq data from mitochondria or whole cells, rather than relying on WGS or WES data, as references. This approach significantly reduces the experimental cost of building the reference data set and the computational cost of identifying the endogenous barcodes. Using a label propagation algorithm, mitoSplitter can reliably demultiplex scRNA-seq data from human cell lines or clinical samples with high accuracy and computational efficiency. This allowed us to analyze 60,000 cells from 10 donors in a single experiment. Moreover, we utilized mitoSplitter to identify a specific synthetic lethality of Bromodomain and extra-terminal (BET) and *TOP2A* inhibition in the non-small cell lung cancer (NSCLC) cell line A549.

Results

Working principle of mitoSplitter

Compared to other chromosomes, the mitochondrial genome is much shorter, with only 16,569 base pairs, making a greater sequencing depth feasible compared to a different chromosome (**Supplementary Figure 1**). The shorter length of the mitochondrial genome significantly reduced the required computational resources. In addition, the number of mitochondria in a cell can vary greatly, with some cells containing more than 2,000 mitochondria, resulting in a rich diversity of mitochondrial genome sequences. All of these characteristics make the mitochondrial genome an ideal genotyping material. To estimate whether the mitochondrial genome variants within a population were diverse enough for genotyping, we analyzed a published RNA sequencing data from Gene Expression Omnibus with accession number GSE107011[17]. We found that the average mitochondrial SNP profile diversities between individuals were approximately 1.5-1.7 times greater than those found in other chromosomes (**Figure 1A-B,** the average mitochondrial SNP profile variations between individuals were also calculated using the whole genome sequencing data from the 1,000 genomes project (Phase 3)[18] detailed in **Supplementary Figure 2**). Therefore, we hypothesized that it is indeed feasible to use mtRNA variants for genotyping based on bulk RNA-seq data of either mitochondria or whole cells. Furthermore, we contend that it is possible to demultiplex the scRNA-seq data from mixed samples (from different donors) based on mtRNA variants.

Subsequently, we established an algorithm named "mitoSplitter" to demonstrate these hypotheses (**Figure 1C**). Briefly, the mitochondrial SNPs were identified through alignment and analysis of the RNA-seq data (including both bulk and single-cell) to the reference genome for each sample. The highly-variable mitochondrial SNPs from the bulk data were selected for mitochondrial genotyping according to the variant's specific frequency and dispersions. Once mitochondrial genotyping of samples was confirmed, a label representing which sample the individual cell was derived from was inferred according to correlation throughout bulk samples and cells, which was then included in the metadata of the scRNA-seq dataset. Due to the low depth of single-cell sequencing, some cells did not receive reads mapped to their selected highly variable sites (mitochondrial genotyping sites) and therefore did not receive labels. To overcome this challenge, a semi-supervised machine learning strategy, known as a label propagation algorithm (LPA), was integrated to transfer sample labels among cells according to specific mtRNA SNPs

detected in single cells. Finally, the distribution of cells from different samples was visualized through the assignment of different colors to the samples on the TSNE plots. The source code and documentation for this computational algorithm, mitoSplitter, are available on Github (https://github.com/lnscan/mitoSplitter).

Validation of mitoSplitter using virtually multiplexed scRNA-seq data

We first tested the reliability and robustness of mitoSplitter using two simulation datasets with different mixing scenarios involving 4 and 3 PBMC samples (detailed in **Supplementary Table 1 and 2**). The genotypes were obtained from the pooled scRNA-seq data for each donor (the sequencing depth of each donor's mitochondrial genome is shown in **Supplementary Figure 3A and B**). We compared the accuracy, running time, and required memory of mitoSplitter to four published computational algorithms, including Demuxlet[9], Souporcell[13], scSplit[11], and Vireo[12]. The results revealed that mitoSplitter demonstrated comparable or even higher accuracy (measured as AUC, see the details in **Supplementary Note 1**) than other tools for both datasets. However, the computation time and memory requirement of mitoSplitter were much lower than those of other methods (**Figure 2A**). Moreover, the accuracy of mitoSplitter's demultiplexing was also validated using a TSNE plot based on mtRNA variants, where the "True" label from the original datasets and the "Predicted" label from mitoSplitter showed strong agreement (**Figure 2B**, the annotated cell types in the demultiplexed data are shown in **Supplementary Figure 4A and B**).

We tested mitoSplitter and the other methods using a large virtually mixed scRNA-seq dataset comprising 60,000 cells from across 10 samples (**Supplementary Table 3**). MitoSplitter demonstrated decent accuracy within 6 hours on a server with ~80 Gigabytes (G) of memory, making it an ideal tool for splitting large datasets. In contrast, scSplit and Vireo indeed failed to analyze this large dataset with maximal memory (600G) on our server (**Figure 2C**). The accuracy of mitoSplitter was further demonstrated in mtRNA variant-based TSNE plots (**Figure 2D**). These experiments on simulated datasets suggest that mitoSplitter may perform label-free scRNA-seq data demultiplexing with minimal computational resources.

Validation of mitoSplitter using pseudo-mixed PBMC samples

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

We randomly selected a 5-donor pseudo-multiplexed dataset to demonstrate the performance details of mitoSplitter. The analysis of mitoSplitter on all 5 donors showed excellent accuracies (**Figure 3B**). To determine if the same biological conclusion can be drawn from the computationally demultiplexed data and experimentally separated data, we compared the downstream analysis results with the "True" and "Predicted" labels. The mtRNA variants-based TSNE plot illustrate the overall agreement of the data with the "True" and "Predicted" labels from mitoSplitter (**Figure 3C-D**). Moreover, the TSNE plots showed the universal distribution of 8 cell types from the 5 donors (**Figure 3E-F**). There was no significant difference in the proportion of each cell type with the "True" and "Predicted" labels in the data from the five donors, as indicated by a paired t-test (**Figure 3G, Supplementary Table 6**). Importantly, gene expression-based TSNE plots revealed remarkable batch effects, emphasizing the importance of performing scRNA-seq analysis with mixed samples (**Figure 3H-I**). These results underscore the ability of mitoSplitter 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 mitoSplitter using experimentally multiplexed scRNA-seq data

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

To further determine the robustness of mitoSplitter using primary samples, 39,106 and 31,973 PBMCs from across 5 donors were analyzed in both separated and pooled fashion in parallel (**Supplementary Table 8**). MitoSplitter completed analysis in under 0.37 hours using 14.4G of RAM. 5,149 ± 1,103 cells were demultiplexed per sample (the tSNE results are shown in **Supplementary Figure 5A**). There was also no significant difference in the proportion of each cell type in the separated sequenced and demultiplexed samples from the five donors as verified by a paired t-test (**Figure 4D,** the cell type compositions are shown in **Supplementary Figure 5B and C**). These results suggest that the demultiplexed results from mitoSplitter were comparable to those of separated analysis.

Subsequently, we tested a large label-free scRNA-seq dataset consisting of 25,223 PBMCs from ten donors (**Supplementary Table 9**). MitoSplitter was able to complete the analysis in under 1 hour while using a negligible amount of RAM (27G). 2,142 ± 922 cells were detected per sample. We observed low batch effects between different donors in the gene expression-based TSNE plot (**Figure 4E**). The proportions of each cell type in the demultiplexed data from the ten donors were consistent with the existing knowledge of PBMC (**Figure 4F**). These findings indicate that mitoSplitter can reduce batch effects through a combination of as many as ten samples into a single scRNA-seq.

178 **scRNA-seq analysis of NSCLC using mitoSplitter identified synthetic lethality of a BET**
179 **degrader and *TOP2A* inhibitor**

180 NSCLC is a significant cancer type worldwide. Previous research has demonstrated that the
181 inhibition or degradation of bromodomain and extra terminal (BET) proteins may impede the
182 proliferation of NSCLC cells[19-21]. However, it has also been reported multiple times that the
183 sensitivity of NSCLC cells to chemical inhibition or degradation of BET is highly variable[22-25].
184 Consequently, the drugs which can induce the synthetic lethality to overcome the resistance to
185 BET inhibitors are still missing. The potential reasons for the inconsistences among these studies
186 include the lack of analysis in single cell resolution as well as the batch effects of different studies.
187 Therefore, we were aiming to understand the mechanism for NSCLC's resistance to BET inhibitor
188 by employing multiplexed single cell RNA-seq and mitoSplitter.

189 Firstly, we observed different effects of chemical degradation of BET through the treatment of
190 ARV-771[26] on the proliferation rate of NSCLC cell lines (**Supplementary Figure 6A**).
191 Interestingly, the A549 cell line exhibited significant resistance to ARV-771, prompting us to
192 investigate the underlying resistance mechanism using mitoSplitter.

193 To mitigate batch effects, we performed multiplexed scRNA-seq analysis of five NSCLC cell
194 lines, treated or untreated with ARV-771 for 48 hours, in a single experiment. The analysis revealed
195 multiple clusters purely based on expression profiles (**Figure 5A**). Using bulk RNA-seq data of
196 mitochondria, we assigned the cells to their origins and presented them in the TSNE plots (**Figure**
197 **5B, Supplementary Table 10**). Notably, ARV-771 treatment resulted in a decrease in the number
198 of cells, except for A549 cells (**Figure 5C**).

199 We investigated the mechanism of resistance to ARV-771 by performing a computational
200 analysis of the scRNA-seq data from A549 cells. We observed the proliferation of a cluster
201 (subgroup 1) in the presence of ARV-771 (**Figure 5D and E**) and identified a list of markers for
202 this cluster (**Supplementary Table 11**). Functional enrichment analysis of these markers revealed
203 their association with mitosis and cell cycle (**Supplementary Figure 6B and C**). Notably, we
204 observed an upregulation of *TOP2A* which is a well-known therapeutic target for NSCLC (**Figure**
205 **5F**). Importantly, the combined chemical inhibition of *TOP2A* with etoposide and ARV-771
206 synergistically inhibited the proliferation of A549 cells specifically (**Figure 5G, Supplementary**
207 **Figure 6D**).

Discussion

In the current study, we present a method called "mitoSplitter" for analyzing scRNA-seq data from multiple samples in a single experiment. This approach uses variants of mtRNA to deconvolute individual samples from pooled scRNA-seq data with reasonable computational time and memory usage. Our application of mitoSplitter to NSCLC revealed a synthetic lethality between *TOP2A* inhibition and BET chemical degradation.

The high cost of scRNA-seq limits its application in studying complex diseases, which often require analysis of a large number of samples. To increase throughput, one strategy is to multiplex multiple samples in a single experiment[27-30]. However, existing strategies using endogenous or exogenous barcodes incur significant experimental or computational costs. In contrast, mitoSplitter only requires additional low-depth sequencing of mtRNA of each sample, which dramatically reduces the overall cost of multiplexed scRNA-seq assays. We anticipate that mitoSplitter will be useful for epidemiology and population genetics studies.

The batch effects among different experiments represent a major technical challenge for scRNA-seq analysis. Despite several efforts from the community[31-35], performing multiplexed scRNA-seq analysis in a single experiment is still the ideal method to overcome this issue. The existing methods of demultiplexing scRNA-seq data using either exogenous or endogenous barcodes require extensive experimental or computational efforts. By launching mitoSplitter, we are now able to demultiplex label-free scRNA-seq data with minimal computational expertise. Using mitoSplitter, biologists with limited experimental and computational capability can analyze mixed samples from scRNA-seq assays.

Before developing mitoSplitter, we were concerned that the lack of diversity in SNPs within the mitochondrial genome may hinder the demultiplexing of complicated samples. However, our study validated the robustness of mitoSplitter using various datasets containing 2-10 samples. Compared to other methods, mitoSplitter exhibited comparable or even higher accuracy in assigning cells to samples, while requiring less RAM and running time for all tested datasets. Notably, mitoSplitter performed similarly well on scRNA-seq data from both cell lines and primary cells.

The BET inhibitor or chemical degrader has shown potential as an anti-cancer agent for a wide range of cancer cells. Therefore, it is crucial to understand the mechanism of resistance to ARV-771 in A549 cells. Single cell RNA-seq is a useful assay for resistance studies since it can identify

239 the features of resistant clones. But due to the batch effects of single cell RNA-seq assay, it is
240 challenging to analyze the data of resistant and sensitive cells generated from separated
241 experiments. In the current study, application of mitoSplitter on multiplexed single cell RNA-seq
242 data revealed a clear clonal evolution event of the resistant A549 cells, when they were mixed
243 other four sensitive cell lines and treated with BET degrader ARV-771. These results led us to
244 further identify a previously unknown synthetic lethality between BET degraders and *TOP2A*
245 inhibitors. However, the detailed mechanism underlying the emergence of the resistant clone
246 requires further investigation.

247 Currently, mitoSplitter is limited by the requirement of mitochondrial genome information in
248 the scRNA-seq data, making it unsuitable for single nucleus RNA-seq (snRNA-seq) analysis. As
249 snRNA-seq is becoming a popular assay for analyzing frozen clinical samples, methods
250 specifically designed for demultiplexing snRNA-seq data will need to be established in the future.

251 In conclusion, we developed the mitoSplitter assay to reliably and efficiently analyze multiple
252 single-cell samples in one scRNA-seq experiment. With mitoSplitter, we have overcome many
253 technical obstacles for biologists to perform scRNA-seq experiments with mixed samples. The
254 availability of mitoSplitter will further accelerate the application of scRNA-seq assay in
255 biomedical research.

256

Material and Methods

Cell preparation

Whole blood samples were obtained from donors with approval from the ethical committee of Xiamen University (Protocol No. KYX-2018-006). To isolate PBMCs, 3 mL of whole blood was fully mixed with 3 mL of 1×PBS and subsequently added onto 6 mL Ficoll-Paque Plus (GE Healthcare, 17-1440-02). After centrifugation at 2000 rpm at 15°C for 30 minutes, the PBMC layer was collected, washed twice with 1×PBS, and resuspended in 300μL of RPMI-1640 (Hyclone, SH30809.01) supplemented with 10% FBS (Viva cell, C04001-500) and 900μL red blood cell lysis buffer (Solarbio, R1010). The sample was then placed on ice for 15min, followed by centrifugation at 2000 rpm at 4°C for 3 minutes. After washing with 1×PBS, cells were resuspended in 1×PBS for the mitoSplitter experiment. Cell viability was confirmed using Countess™ Cell Counting Chamber Slides (Invitrogen, C10283).

H2228, HCC827, H1975, A549, and H838 cells were cultured in RPMI-1640 supplemented with 10% FBS and 1% penicillin-streptomycin (Invitrogen, 15140122) at 37°C and 5% CO_2. For scRNA-seq samples, the five cell lines were co-cultured with equal cell numbers in two culture plates, one with 0.2 μM ARV-771 (MedChemExpress, HY-100972) treatment for 48 hours and another without drug treatment. For bulk mtRNA samples, cells from the five cell lines were cultured independently without drug treatment. To collect cells, the media was removed, and cells were washed with 1×PBS before detachment from the culture plates with Trypsin (Gibco, 25200072) and incubated at room temperature for 2 minutes. After deadhesion, cells were harvested with complete media, washed three times with 1×PBS, and resuspended in 1×PBS for the mitoSplitter experiment.

Sequencing library preparation for mitoSplitter

The Cell Mitochondria Isolation Kit (Beyotime, C3601) was used to isolate mitochondria from individual samples for bulk mitochondrial RNA (mtRNA) sequencing, following the manufacturer's instructions. RNA was then purified from the isolated mitochondria using the Total DNA/RNA Isolation Kit (OMEGA, R6731-01). The library preparation for bulk mtRNA sequencing followed a modified Smart-Seq2 protocol[36,37]. In brief, purified mitochondrial RNA (5 μL) was subject to reverse transcription by adding 2 μL 10 μM oligo-dT (ordered from Sangon, **Supplementary Table 12**), 2 μL 10 mM dNTPs (TransGen Biotech, AD101-11), 4 μL 5×RT buffer, 1 μL Maxima H Minus reverse transcriptase (Thermo Scientific, EP0752), 1 μL 50 μM Template

Switch oligo (ordered from Sangon, **Supplementary Table 12**), 1 μL RNase Inhibitor (TransGen Biotech, AI101-02), 2 μL 10 mM GTP (Invitrogen, R0461), and 2 μL 50% PEG-8000 (Beyotime, R0056). The reverse transcription was then incubated at 42°C for 90 minutes followed by 5 cycles of 50°C for 2 minutes and 42°C for 2 minutes, and a final incubation at 70°C for 15 minutes. For PCR, 25 μL 2×HiFi HotStart Readymix (Kapa Biosystems, KK2602), 0.8 μL 50 μM ISPCR oligo (ordered from Sangon, **Supplementary Table 12**), and 4.2 μL Nuclease-free water were added. The PCR reaction was incubated at 98°C for 3 minutes followed by 10 cycles of 98°C for 20 seconds, 67°C for 15 seconds, and 72°C for 6 minutes, and a final incubation at 72°C for 6 minutes. The procedure above was independently performed on cells from different donors or cell lines.

For single-cell RNA sequencing, the chip fabrication, barcode bead preparation, and sequencing library preparation were performed as previously described[18]. To prepare pseudo-mixed PBMC samples, cells from 11 donors were subject to scRNA-seq independently, and 10,000-dual-well chips were used. To prepare experimentally multiplexed scRNA-seq data, PBMCs from 5 donors were subject to independent and multiplexed scRNA-seq (cells from different donors were mixed thoroughly with the equal input before loading to the Well-paired-seq chip) simultaneously, and 30,000-dual-well chips were used. To generate a large scRNA-seq dataset from 10 donors, PBMCs from 10 donors were subject to multiplexed scRNA-seq, and 100,000-dual-well chips were used for scRNA-seq.

The amplification products of both bulk and scRNA-seq samples were purified twice using 0.6× VAHTS DNA Clean Beads (Vazyme, N411-02), and cDNA concentrations were measured with a Qubit fluorometer. An appropriate volume of cDNA was used as an input for library construction using the TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme, cat#TD502). The reaction was performed following the instructions of the manufacturer except for the primers. For bulk sample and scRNA-seq sample preparation, primer N501 and primer P5-TSO_Hybrid (**Supplementary Table 12**) were used as the P5 primer, respectively. Additionally, primers Nextera_N7×× (**Supplementary Table 12**) were used as the P7 primer in place of the kit's provided oligonucleotides. The libraries were purified twice using 0.6× VAHTS DNA Clean Beads, and cDNA concentration was measured again. Finally, the libraries of bulk samples and scRNA-seq samples were sequenced on Illumina Novaseq 6000 and HiSeq X Ten instruments with paired-end 150 base reads, respectively.

318 **Identification of different cell proliferation rates in NSCLC cell lines under ARV-771**
319 **treatment**
320 The H2228, HCC827, H1975, A549, and H838 cell lines were cultured, harvested
321 independently, and diluted to 5 cells/μL using complete culture media. Subsequently, 100 μL of
322 cell suspension for each cell line was added to different wells of 96-well plates. After a 24-hour
323 culture period for cell adhesion, each cell line was exposed to ARV-771 at six different doses
324 (0.0625 μM, 0.125 μM, 0.25 μM, 0.5 μM, 1 μM, and 2 μM) in six replicates. The experiment
325 included a total of 42 independently treated cell populations of each cell line, including a vehicle
326 control. After a 48-hour treatment, the cell viability of each condition was measured using the cell
327 counting kit-8 assay (Apexbio, K1018-5). The IC_{50} values of each cell line were calculated to
328 determine the best ARV-771 concentration to result in different cell proliferation rates in NSCLC
329 cell lines.

330 **Validation of the effect of combined ARV-771 and etoposide on the proliferation of A549 cells**
331 Pre-experiments were conducted to determine the dose range of etoposide that could
332 effectively inhibit the proliferation rate of A549 cells. Subsequently, A549 cells were exposed to
333 ARV-771 and etoposide (MedChemExpress, HY-13629) at the following combinations of doses,
334 0 μM, 0.0625 μM, 0.125 μM, 0.25 μM, 0.5 μM, 1 μM, and 2 μM of ARV-771, and 0 μM, 0.01 μM,
335 0.1 μM, 0.2 μM, 0.4 μM, 0.8 μM, and 1 μM of etoposide. This experiment included a total of 48
336 independently treated cell populations, including vehicle controls, and each combination had six
337 replicates. Cell viability was measured using the CCK-8 assay, and the combination of 0.5 μM
338 ARV-771 and 0.2 μM etoposide was selected for statistical analysis with controls.

339 **Data pre-processing**
340 To verify the reliability and robustness of mitoSplitter, we collected data from several publicly
341 available single-cell sequencing datasets.
342 General single-cell sequencing data included three datasets. Dataset 1 was obtained from Gene
343 Expression Omnibus (GEO) with accession number GSE96583[9] and consisted of 3,177 PBMCs
344 from 4 individuals (detailed in **Supplementary Table 1**). Dataset 2 was obtained from GSE213096
345 and consisted of 25,078 PBMCs from 3 individuals (detailed in **Supplementary Table 2**). Dataset
346 3 was obtained from the GSA-Human repository with accession number HRA001748 and
347 consisted of 60,000 liver cells from 10 individuals (detailed in **Supplementary Table 3**).

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

Four sets of sequencing data, including mitochondrial bulk RNA-seq generated using a modified Smart-Seq2 protocol[36,37], and single-cell RNA-seq data generated from the Well-paired-seq[18] platform, were generated. The first set covered 42,779 PBMCs from 11 individuals (dataset 4, detailed in **Supplementary Table 4**). The second set consisted of PBMCs from 5 donors and included 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 was a multiplexed single-cell RNA-seq dataset 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 experiments 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 matrices

To generate single-cell and bulk mitochondrial mutation matrices, three steps were undertaken: sequence alignment, remapping, and variant calling. The scRNA-seq data were aligned to the GRCh38 genome using CellRanger v3.1.0 (for 10x Genomic RNA-seq data) or zUMIs v2.9.7e[39] (for Well-Paired-Seq data), during the sequence alignment process, and the bulk mitochondria RNA-seq data were quality-filtered using cutadapt v1.18[40] and aligned to the mitochondrial sequence in the GRCh38 genome using STAR v2.7.3a[41]. To improve the quality of alignment and reduce the false positive rate of variant calling, the obtained bam files were remapped to the reference genome using minimap2 v2.24[42] with 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. The variant calling was performed using a python mtRNA genotyping pipeline[16]. In detail, reads in the remapping single-cell bam file with base quality or alignment quality lower than 20 were filtered, and the allele frequency of each cell was then extracted. For the bulk

379 mitochondria RNA-seq bam file, the threshold value of base quality and alignment quality was set
380 to 10. Finally, all the mutations from all samples or cells were integrated to create a bulk or single-
381 cell mutation matrix.

382 **Demultiplexing of mixed sequencing data using mitoSplitter**

383 We used scanpy v1.9.1[43] to select high-variable SNPs across bulk samples from a
384 mitochondrial mutation matrix for mitochondrial genotyping. These mitochondrial sites were more
385 donor-specific, so we calculated the correlation between single cells and each bulk sample based
386 on the variants frequency of these sites to evaluate the probability of cells from each donor. A
387 single cell was considered to be from the same donor as the bulk sample whose bulk sample
388 exhibited the highest correlation unless it lacked a positively correlated bulk sample. Cells that
389 lacked mitochondrial genotyping sites were unable to receive donor labels through this pipeline
390 due to the uneven coverage of scRNA-seq data. To infer the donor origin of each unbiased cell, we
391 predicted it based on the mtRNA SNPs in scRNA-seq and the inferred labels. We defined this task
392 as a semi-supervised learning problem and used LPA to propagate labels through cells using
393 mtRNA SNPs in scRNA-seq.

394 In detail, we classified (X, Y) into labeled data as $(x_1, y_1) \ldots (x_l, y_l)$ and unlabeled data as $(x_{l+1},$
395 $y_{l+1}) \ldots (x_{l+u}, y_{l+u})$, depending on whether the sample source labels were obtained. X was the matrix
396 of mtRNA SNPs, and Y was the matrix of cell sample source label, $\mathbf{Y_L}=\{y_1, \ldots y_l\}$ consisted of
397 inferred labels and $\mathbf{Y_U}=\{y_{l+1} \ldots y_{l+u}\}$ was undetermined. The task was to estimate $\mathbf{Y_U}$ from $\mathbf{X}$ and
398 $\mathbf{Y_L}$. LPA created a fully connected graph according to the similarity matrix constructed from $\mathbf{X}$,
399 where the nodes represented all cells. The edges between nodes were weighted by calculating the
400 normalized graph Laplacian matrix. Then, labeled nodes acted as sources to spread labels $\mathbf{Y_L}$
401 among cells through the edges. Larger edge weights meant easier label propagation. Finally,
402 unlabeled cells received persistent sample labels $\mathbf{Y_U}$ when the network of the graph reached a
403 globally stable state. We set the clamping factor α=0.8, which helped the algorithm keep 80% of
404 input labels and change the confidence of labels within 20% during propagation. This optimization
405 was based on inferred labels from unlabeled data. Based on the LPA model, we calculated the
406 probability of each unlabeled cell belonging to each donor. The final cell sample label was
407 determined by the LPA result with a probability greater than $(1\text{-}10^{-10})$.

408

Data dimension reduction and visualization

The effect of mitochondrial mutation on demultiplexing was illustrated using mtRNA mutation-based tSNE plots. Similar to the workflow for single-cell gene expression, Seurat was used to normalize data, as well as scaling, clustering, and dimension reduction based on 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 utilized a combination of multiple single-cell sequencing datasets. Based on the number of mixed samples k and the number of singlets predicted by scrublet v0.2.3[49], we randomly selected the specified number of cells presented within each sample followed by the random selection of $1/k$ of the reads from their corresponding reads for mixing. To remove duplicated reads with the same read ID from different samples, we used the seqkit v2.3.0[44] rmdup function. The remaining reads were mapped to the GRCh38 genome using CellRanger v3.1.0 or zUMIs v2.9.7e[39]. After remapping and variant calling, a single-cell mtRNA mutation matrix could be obtained. Finally, the bulk mtRNA mutation matrix was produced by combining the reads of all cells within a sample.

Computational benchmarking.

We compared four demultiplexing algorithms with mitoSplitter: Souporcell v2.0[13], demuxlet v2[9], scSplit v1.0.9[11], and Vireo v0.2.3[12]. Each algorithm utilized a maximum of 50 threads, and default pipeline and parameters were used for all demultiplexing calculations. For Souporcell, we ran souporcell_pipeline.py through the singularity container. For demuxlet, we mapped single-cell RNA-seq data from each sample to the GRCh38 genome using CellRanger v3.1.0, generated bam files for genotype calling, and output genotypes through VCF files by samtools v1.9[45] and bcftools v1.8[45]. We merged the VCF files into a single file using VCFtools v0.1.17[46], and obtained the demultiplexing results using popscle v0.1-beta demuxlet function, with merged VCF file and "GT" field. For scSplit, we conducted quality-filtering 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'". We generated the scSplit count matrix using the scSplit count function and conducted demultiplexing using scSplit run function with "-d 0".

To conduct performance verification on pooled single-cell sequencing data, we calculated the AUC value, accuracy, and recall of single cells using the Python package sklearn v1.0.2. Singlets predicted by scrublet v0.2.3[49] in each sample were used for performance evaluation. We recorded

the running time and memory requirements of the algorithms using custom scripts.

Cell clustering and annotation

We used Seurat v4.3.0[50] 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. Cell type annotation was conducted using celltypist v1.3.0[51] with the "Immune_All_Low.pkl" model. We annotated each Seurat cluster with the top one or two cell types by proportion, and identified marker genes for each cell type using the "FindConservedMarkers" function in Seurat. For each cell type, we used a paired t-test to compare cell type proportions across samples based on the true and predicted labels.

Data availability

The datasets used in the present study are all publicly available. Dataset 1 was obtained from Gene Expression Omnibus with accession number GSE96583. Dataset 2 was obtained from GSE213096. Dataset 3 was obtained from the GSA-Human repository with accession number HRA001748. Single-cell RNA-seq data with exogenous barcodes were downloaded from European Nucleotide Archive repository, under accession number E-MTAB-9964 for CellPlex and TotalSeq-B, and E-MTAB-11401 for LMO (custom) and TotalSeq-A. Raw Fastq files in dataset 4-6 and non-small cell lung cancer cell lines experiments will be deposited after the article is received.

Code availability

The code is available on https://github.com/lnscan/mitoSplitter.

Acknowledgments

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., R.C., X.W., G.Y., Z.W., Y.Z.

Conflict of Interest Statements

The authors claim no conflict of interest.

References

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).

Figures

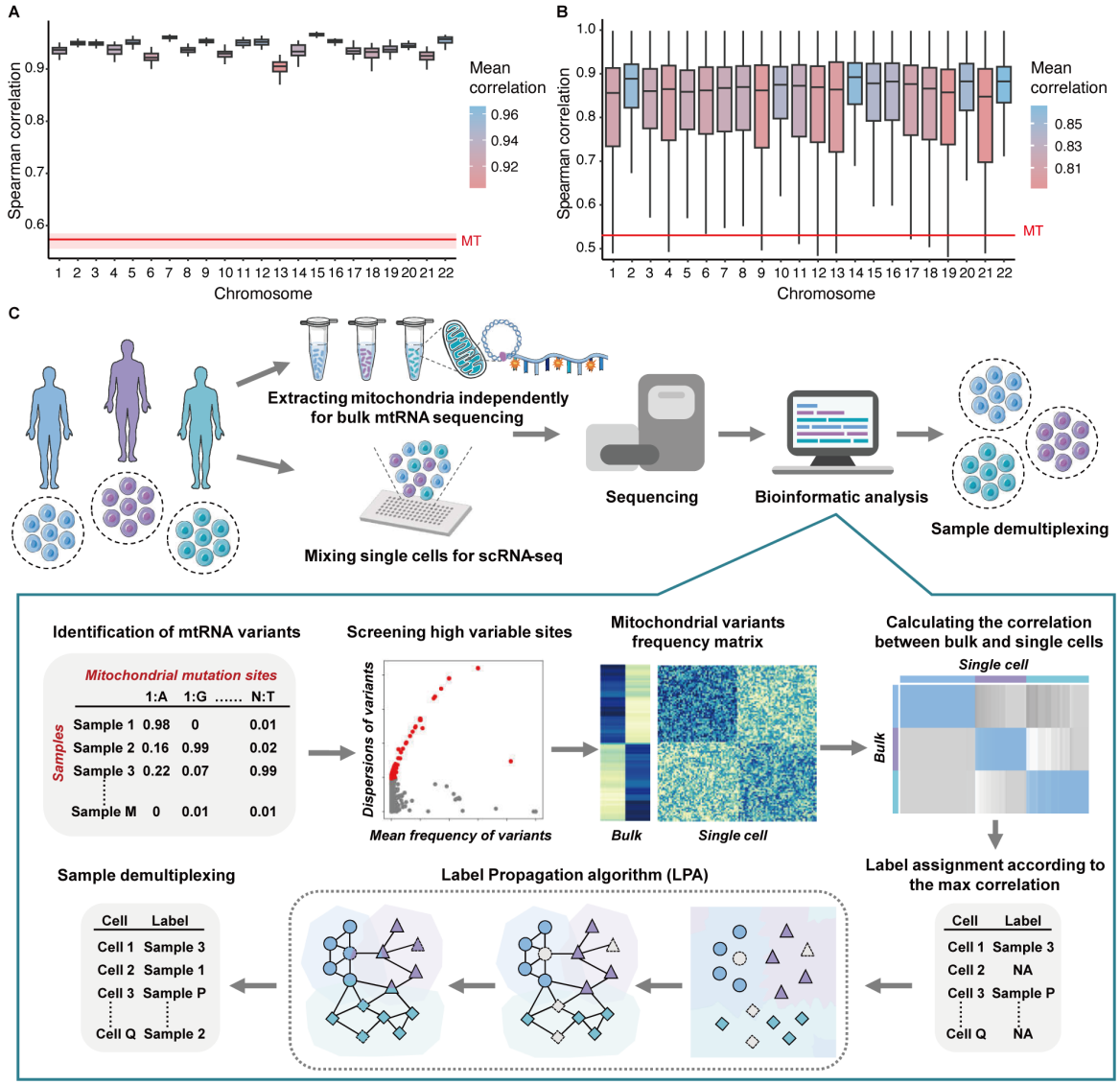

Figure 1. Mitochondrial variants as the optimal material for genotyping. (A) The mean Spearman correlation between all pairs of two individuals in a published RNA sequencing data based on mitochondrial or autosomal variant profiles. The red line represents the mean correlation of the mitochondrial variants between individuals, while the upper and lower pink boundaries represent the first and third quantiles. MT represents "mitochondrion". These results demonstrated that mitochondrial variant profiles are significantly more variable between individuals than those of other chromosomes. (B) The mean Spearman correlation between pairwise individuals from a published RNA sequencing data based on variant profiles of mitochondrial or autosomal segments. An autosomal segment is composed of 13 adjacently located genes (equal to the gene number of the mitochondrial genome), and 100 segments were randomly extracted from each autosomal segment. The red line represents the mean correlation of the mitochondrial genome between individuals. MT represents "mitochondrion". These results indicate that profiles of mitochondrial variants are significantly more variable

604 than those of other chromosome segments containing the same number of genes. (C) Overview of mitoSplitter
605 framework. SNPs were identified by aligning RNA-seq data to the reference genome for each sample. High-
606 variability mitochondrial SNPs were identified based on variation frequency and dispersions in bulk RNA-seq,
607 and used to genotype samples. The mitochondrial genotyping correlation between bulk samples and cells was
608 used to determine the sample label of some single cells. Only cells with high bulk sample correction were labeled.
609 Then, the label propagation algorithm (LPA) was employed to give sample labels to unlabeled cells based on
610 single-cell mtRNA SNPs.

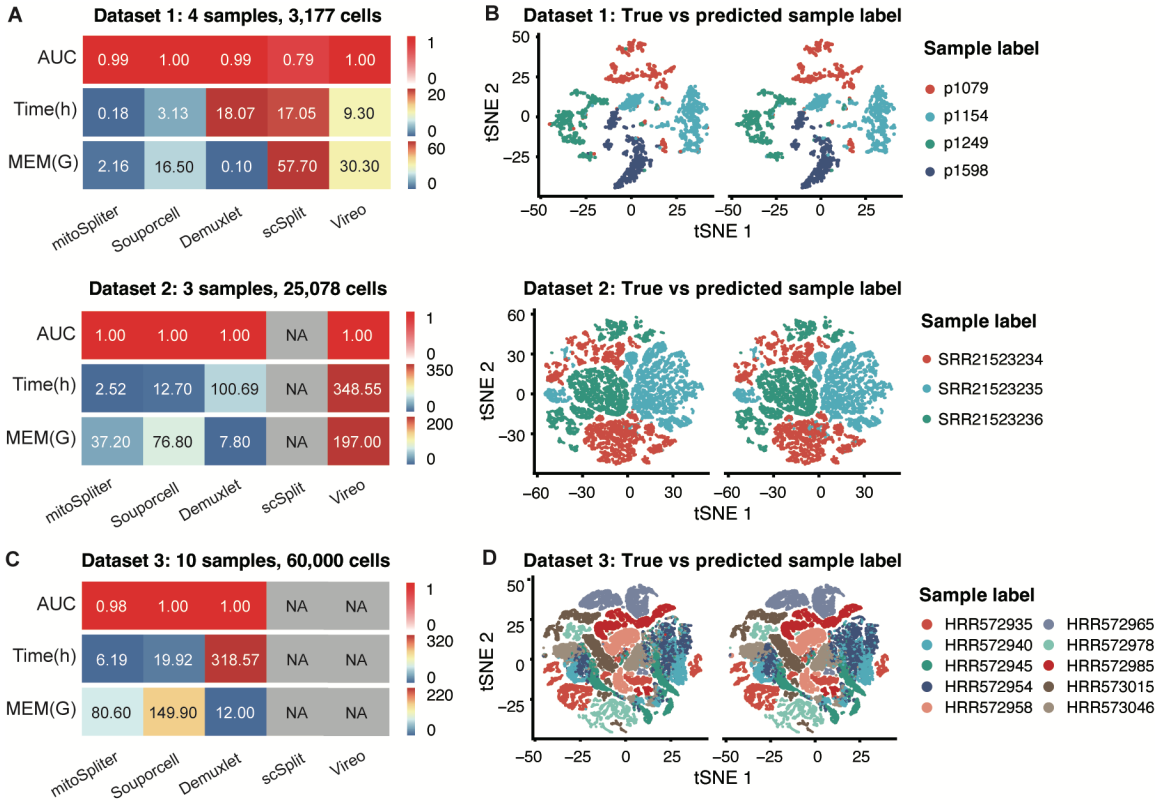

Figure 2. Performance of mitoSplitter on virtually multiplexed scRNA-seq data. (A) Comparison of the performance of mitoSplitter and other demultiplexing tools on two datasets with 4 and 3 samples, displaying AUC (area under the receiver operating characteristic curve), running time, and memory requirements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

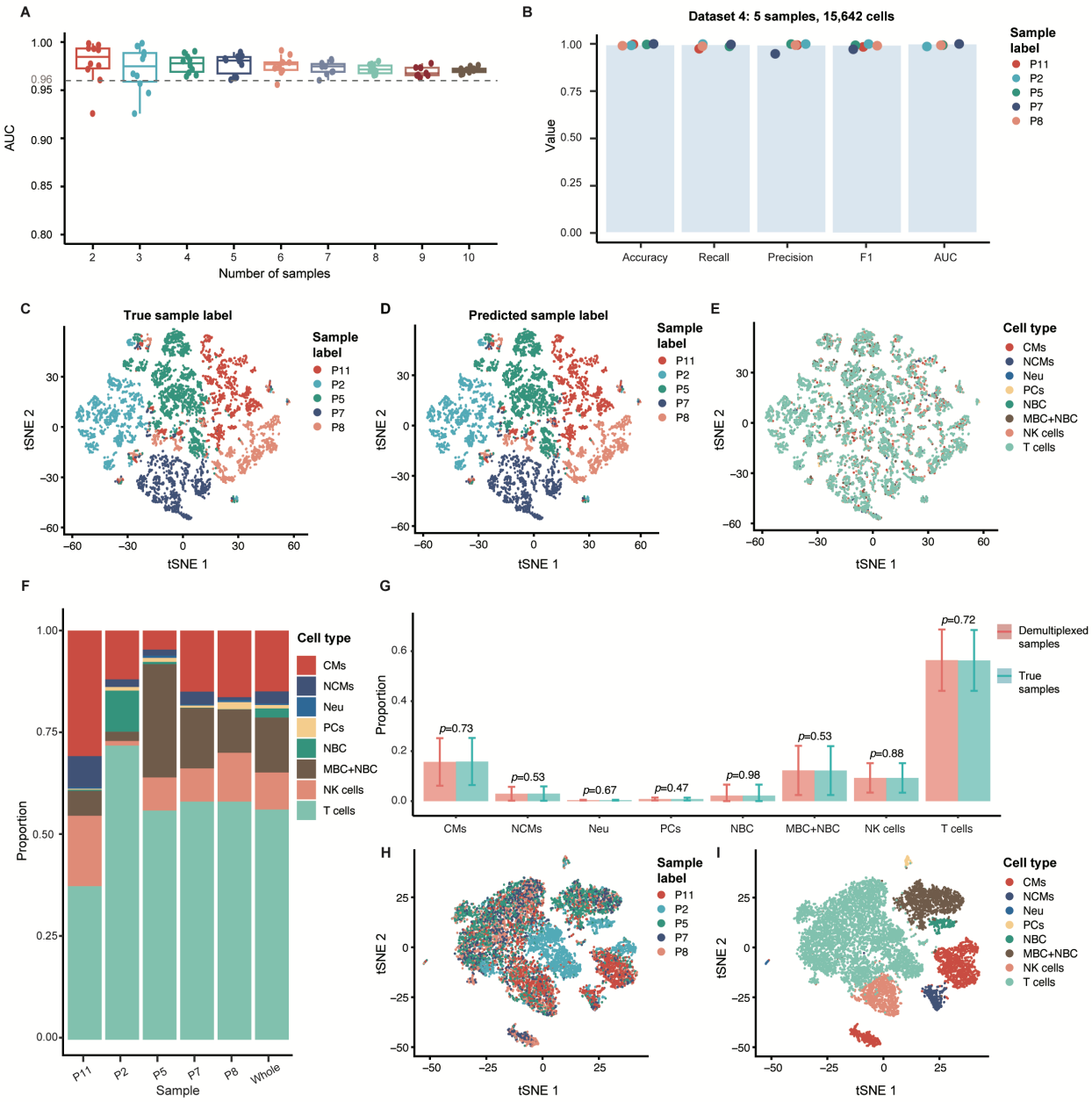

Figure 3. Performance of mitoSplitter on pseudo-mixed PBMC samples. (A) AUC values of mitoSplitter demultiplexing outcomes for varying numbers of mixed samples. (B) Comprehensive evaluation of mitoSplitter demultiplexing results for a dataset of five mixed samples, including demultiplexing accuracy, recall rate, precision, *F*1 score, and AUC for each sample. (C-E) TSNE plot generated using highly variable variants identified by mitoSplitter and colored based on true sample labels, predicted sample labels, and cell types, respectively. (F) Fraction of each cell type in the demultiplexed samples. (G) Comparison of cell type composition between demultiplexed and true samples using a paired t-test. (H-I) Gene expression-based TSNE plots colored based on predicted sample labels and cell types, respectively. Cell types include classical

monocytes (CMs), non-classical monocytes (NCMs), neutrophils (Neu), plasma cells (PCs), naïve B cells (NBC), memory B cells (MBC), and, natural killer cells (NK cells).

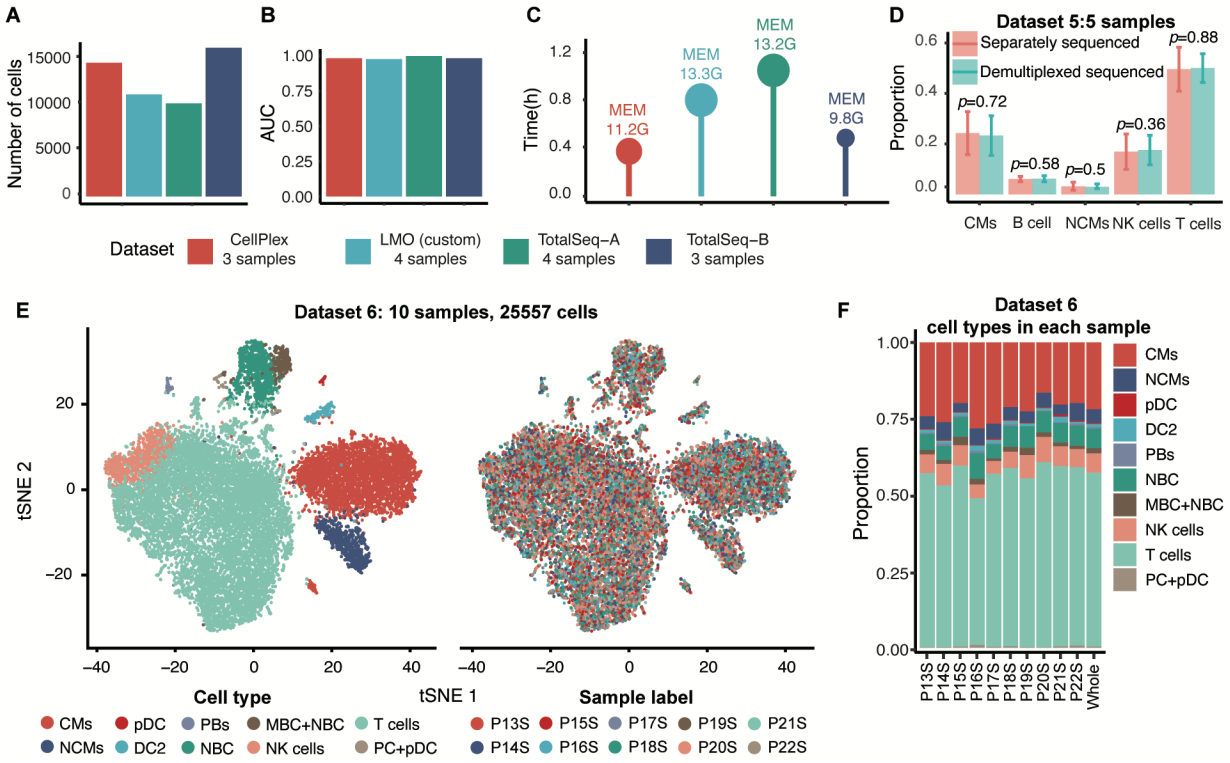

Figure 4. Performance of mitoSplitter on experimentally multiplexed scRNA-seq data. (A) The number of cells in each dataset. (B) The AUC of mitoSplitter on four multiplexed datasets with exogenous barcodes. (C) The running time and memory requirement of mitoSplitter on the same four multiplexed datasets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

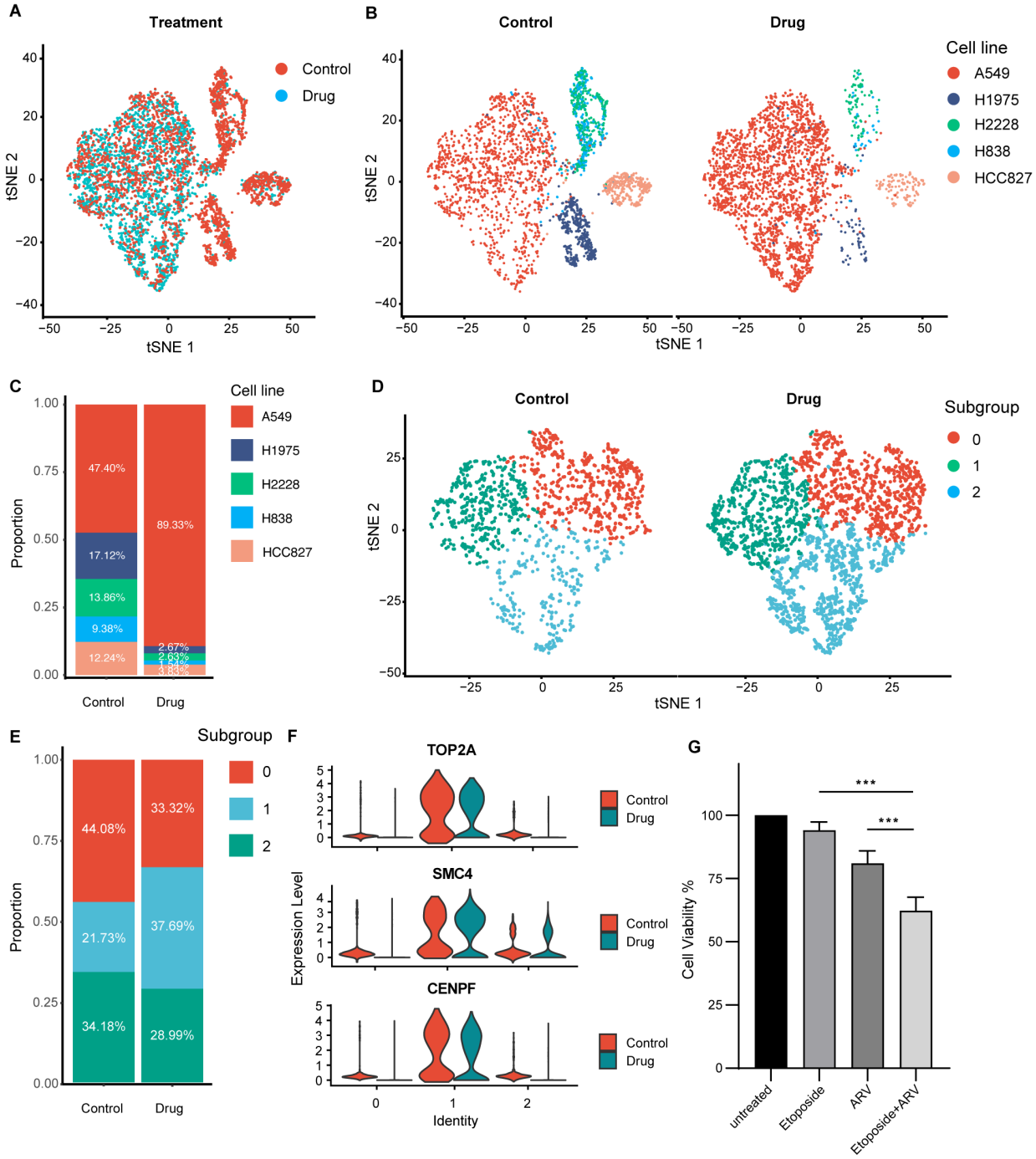

Figure 5. scRNA-seq analysis of NSCLC using mitoSplitter identified synthetic lethality of a BET degrader and *TOP2A* inhibitor. (A) The NSCLC cells are presented in a combined gene expression-based TSNE plot, colored by drug treatment status. (B) Two separate TSNE plots of NSCLC cells based on gene expression, colored by cell lines. The left plot shows cells that were not treated with the drug, while the right plot shows cells that were treated. (C) The proportion of cell lines in the sequencing results before and after drug treatment. (D) Two gene expression-based TSNE plots of A549 cells, where the left plot shows cells that were not treated

with the drug, while the right pot shows cells that were treated. Cells are colored by subgroup types. (E) The proportion of subgroups in the A549 cell line before and after drug treatment, where the proportion of cells belonging to subgroup 1 increased after drug administration. (F) The expression of the top three up-regulated genes identified in A549 subgroup 1 across all subgroups. (G) Combined administration of Etoposide and ARV can significantly reduce the viability of A549 cells.