**Title:** Expression Based Species Deconvolution and Realignment Removes Misalignent Error in Multi-species Single Cell Data

**Authors:** Jaeyong Choi[1,2], Woochan Lee[1,2], Jung-Ki Yoon[3] and Jong-Il Kim[1,2]

**Affiliations:** [1]Department of Biomedical Sciences, Seoul National University, Seoul, Republic of Korea. [2]Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul, Republic of Korea. [3]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

**Correspondence:**

Jong-il Kim, MD, PhD

Department of Biomedical Sciences

Seoul National University College of Medicine

103, Daehak-ro, Jongno-gu, Seoul 03080, Korea

Tel: +82-2-740-8004

Fax: +82-2-742-5947

Email: jongil@snu.ac.kr

***Abstract***

Background

Although single cell RNAseq of xenograft samples are widely used, there is no comprehensive pipeline for human and mouse mixed single cell analysis.

Method

We used public data to assess misalignment error when using human and mouse combined reference, and generated a pipeline based on expression-based species deconvolution with species matching reference realignment to remove errors. We also found false-positive signals presumed to originate from ambient RNA of the other species, and use computational method to adequately remove them.

Result

Misaligned reads account to on average 0.5% of total reads but expression of few genees were greatly affected leading to 99.8% loss in expression. Human and mouse mixed single cell data analyzed by our pipeline clustered well with unmixed data. We also applied our pipeline to multi-species multi-sample single cell library containing breast cancer xenograft tissue and successfully identified all identities along with the diverse cell types of tumor microenvironment.

Conclusion

We present our pipeline for mixed human and mose single cell data which can also be applied to pooled libraries to obtain cost effective single cell data. We also address consideration points when analyzing mixed single cell data for future development.

Keyword: single cell sequencing, patient derived xenograft, bioinformatics pipeline

## *Introduction*

Single cell RNA sequencing is a powerful method which generates transcriptome with greater resolution[1], identify rare cells[2], and compare cell heterogeneity[3]. Single cell technology also provides additional functionality as to discriminate expression of human cells from mouse cells in human-mouse xenograft specimens or human-mouse cell co-culture data[4,5]. A simple method to process human and mouse mixed single cell data is to align the data to human and mouse combined reference (henceforth called "combined reference"). CellRanger from 10x Genomics provides a combined reference and also flags barcodes where both human and mouse cells are presumably contained in a same droplet (henceforth called "cross-species doublet") using read count distribution. But similar sequences between each reference may cause human originated reads to align to mouse reference[6,7]. Also as many of mouse and human genes are homologous, some reads may have best matching sequence in both references, leading to multi-mapped reads.

We measure the extent of misalignment error in single cell data when using combined reference to align human and mouse mixed single cell data and present a expression-based species deconvolution with species matching reference realignment pipeline to minimize error in downstream analysis. We expand the concept and apply the pipeline to multi-sample, multi-species pooled single cell library which can be used to generate cost-effective single cell libraries while removing sequencing batch effects.

## *Method*

### *Tissue preparation*

Human Lung tissue was obtained from lung lobectomy surgery for lung cancer. To get a cancer-free sample, the pathologist confirmed the absence of cancer cells. Human lung organoid was derived from lung of the transplantee undergoing lung transplantation. The Seoul National University Hospital IRB granted permission for this study (2008-065-1148). Patient-derived xenograft from breast cancer patient was sacrificed six weeks after injection. The PDX model for this study was approved by The Seoul National University Hospital IRB (1805-151-948). All study was conducted in accordance with the Declaration of Helsinki and the Guidelines for Good Clinical Practice and we obtained informed consent from each patient

Fresh human lung tissue was treated with Multi Tissue Dissociation Kit 1 (Miltenyi Biotec) through gentleMACS Octo Dissociator with Multi_A program. Resected PDX tumor was dissociated to a single cell with Tumor Dissociation Kit 1, human (Miltenyi Biotec) with TDK_1 program. Both digested tissue suspensions were strained over a 70-um filter. The suspension was centrifuged at 400g for 5min at

4 C with 10ml Advanced DMEM/F12 (Gibco) with Glutamax (Gibco), HEPES (Gibco), Penicillin-Streptomycin (Gibco) (ADF+++). The pellet was resuspended in 1ml 1x Red Blood Cell Lysis Solution (Miltenyi Biotec) for 10min at room temperature for lysis erythrocytes before the addition of ADF+++ and centrifugation at 400g for 5min. The pellets were resuspended in an appropriate volume of ADF+++ and keep on ice for GEM generation.

*Single cell RNA sequencing*

We performed Single cell RNA sequencing according to a standard 10x Genomics 3′ V3.1 chemistry protocol. Before loading the cells on the G chip, cells were counted to determine cell concentration. The number of loading cells per library was calculated via the Multiplexing cost calculator (satijalab.org/costpercell). The libraries were sequenced with Illumina NovaSeq 6000 platform S4 with paired-end 100 bp.

*Public data download*

Human decidua (D12)[8], placenta (D121), liver (P1)[9], kidney (F41)[10], and mouse colon (HC1)[11] single cell data were downloaded from Human Cell Atlas data portal (https://data.humancellatlas.org). Human blood, Hodgkin lymphoma, pooled cancer and mouse brain data were downloaded from 10x Genomics single cell gene expression datasets (support.10xgenomics.com/single-cell-gene-expression).

*Reference comparison*

For each public data, we aligned the data to species matching reference (refdata-gex-GRCh38-2020-A or refdata-gex-mm10-2020-A) using Cellranger 4.0.0, and filtered barcodes by low total read count, low expressed gene count, and high mitochondrial gene percentage. We extracted these barcodes from data aligned to combined reference (refdata-gex-GRCh38-and-mm10-2020-A) and compared the alignment result. We defined misaligned read as 1) read aligned to gene of the other species, or 2) read difference between combined reference and species matching reference. We defined misaligned gene as 1) average read count higher than 0.1 per cell, and 2) more than 10% difference in read count between species matching reference and combined reference. Pathway analysis for misaligned genes were performed with g:Profiler[12].

*Alignment and pre-filter*

We aligned human and mouse mixed single cell data to combined reference using CellRanger 4.0.0. Seurat v3.2.3[13] was used to filter cells, normalize expression, cluster, and visualize data. EmptyDrops[14] was used to identify droplets which had similar expression pattern to ambient RNA. We selected barcodes with expressed gene count >=200 in either human reference and/or mouse reference. We removed barcodes if the percentage of mitochondrial gene were twice or higher than the percent

of mitochondrial gene in ambient RNA for human and mouse genes separately.

*Expression based cross-species doublet identification*

For data containing non-tumor cells, we used scHCL[15] and scMCA[16] correlation value to estimate cross-species doublet. The cutoff was set by normalized correlation value of median absolute deviation >= 2.58 for human and mouse. Barcodes with both high correlation to human cell and mouse cell were classified as cross-species doublet. Other barcodes were classified as human cells or mouse cells according to read majority and correlation value. We removed barcodes with discrepant read majority and correlation value. When expression for each species could be clearly separated by cluster, minor species in each cluster were removed.

For data containing tumor cells, we used copy number cluster result from copyKAT[17] in additional to correlation values to identify tumor cells with low correlation value. We defined aneuploid barcodes as tumor cells. Barcodes with valid copy number result and read majority in mouse were additionally classified as cross-species doublet.

*Species matching reference alignment*

After species identification, we realigned data to human reference and mouse reference, and selected species matching barcodes. When multiple samples were pooled, we used Souporcell[18] to find identities and find inter-sample doublets. We selected barcodes called as singlet by Souporcell, and when expression for each sample could be clearly separated, minor identities in each cluster were removed. Normal cells with noisy copy number segments were also removed.

For each sample, we applied quality filter separately to account for sample variation. cells with low total read count, low expressed gene count, and high mitochondrial gene percentage were filtered. Separately filtered barcodes were merged and DoubletFinder[19] was used to identify intra-sample doublets.

*Ambient RNA removal*

SoupX[20] was used to estimate ambient RNA from barcodes with 100 or less reads. We calculated the contamination rate in human barcodes and mouse barcodes separately to account for ambient RNA from both species. We additionally added 0.05 to the contamination rate to aggressively remove ambient RNA signals. Additional information regarding bioinformatics pipeline can be found at our homepage.

*Data availability*

Mixed and separately sequenced single cell data are available at Sequence Read Archive under number PRJNA719675. Xenograft and lung organoid single cell data are not publicly available due to IRB restrictions and are available from the corresponding author on reasonable request.

### ***Result***

| | | | | matching reference | combined reference | | |
|---|---|---|---|---|---|---|---|
| data name | species | tissue | cell count | total read | read aligned to correct reference | read misaligned to other reference | misaligned read% |
| placentaD121 | human | placenta | 2,522 | 35,877,521 | 35,828,222 | 11,252 | 0.17 |
| liverP1 | human | liver | 957 | 3,179,327 | 3,165,784 | 3,243 | 0.53 |
| deciduaD12 | human | decidua | 1,963 | 25,147,111 | 25,122,973 | 6,205 | 0.12 |
| 10x_pbmc | human | PBMC | 9,295 | 93,171,679 | 93,027,903 | 175,980 | 0.34 |
| kidneyF21 | human | kidney | 3,399 | 21,326,290 | 21,266,741 | 77,660 | 0.64 |
| 10x_hodgkin | human | hodgkin_lymphoma | 2,502 | 19,341,273 | 19,305,209 | 22,330 | 0.30 |
| 10x_SCC | human | SCC | 4,739 | 72,819,855 | 72,583,115 | 482,564 | 0.99 |
| 10x_lungcancer | human | lung_cancer | 4,481 | 50,316,076 | 50,165,228 | 253,394 | 0.80 |
| 10x_adenocarcinoma | human | adenocarcinoma | 3,599 | 44,820,075 | 44,678,244 | 268,975 | 0.92 |
| 10x_NSCLC | human | NSCLC | 6,640 | 69,706,720 | 69,466,298 | 478,684 | 1.03 |
| 10x_mbrain | mouse | brain | 8,096 | 103,833,090 | 103,524,395 | 38,599 | 0.34 |
| mcolonHC1 | mouse | colon | 1,204 | 8,810,265 | 8,782,774 | 2,440 | 0.34 |

Table1: Misaligned reads in combined reference

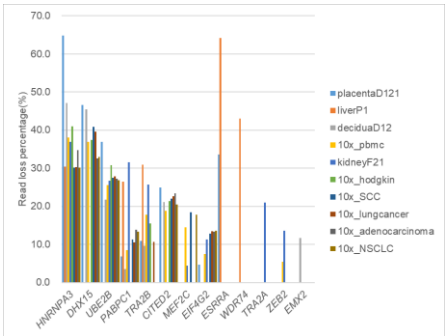



Figure1: Read loss percentage of top misaligned genes

### ***Misalignment in combined reference***

To summarize the extent of misalignment in single cell data, we aligned several public data to combined reference and compared the results to human reference. When aligning human derived single cell data to combined reference, one would expect reads to align to human part of the combined reference. Most reads indeed aligned to human genes but depending on the data type and sequencing throughput, 3000 ~ 480000 reads misaligned to mouse genes. More importantly, when comparing data analyzed with combined reference to human reference, reads aligned to human genes were 13000 ~ 300000 fewer in combined reference. Misalignment error were more prominent in tumor samples, and error were also found when aligning mouse derived single cell data to combined reference. Although all incorrectly aligned reads account to only 0.3 ~ 1% of total reads, these errors focused around small number of genes, leading to strong false signals. Namely 30 ~ 65% of reads of *HNRNPA3* were lost in human cells, and nearly all reads (99.8%) of *Erf1* were

lost in mouse cells. Genes related to RNA binding and HLA genes had high rates of misalignment errors. Interestingly, genes affected by misalignment and the degree of misalignment were not identical between samples and tissues. Genes where reads misalign to and genes where read loss was found did not overlap, suggesting that misalignment was not a simple switch-of-position.

### *Mixed data analysis pipeline*

To assess the misalignment error in real human and mouse mixed single cell data, we generated both mixed and separate libraries of human lung tissue, human gastric organoid, and mouse intestinal organoid. Our mixed data also had misalignment when we aligned the data to combined reference. The error was more intense, possibly owing to ambient RNA contamination and undetected doublets. To overcome this issue, we generated a pipeline based on expression-based species deconvolution with species matching reference realignment to remove doublets and ambient RNA signals for error-free downstream analysis.

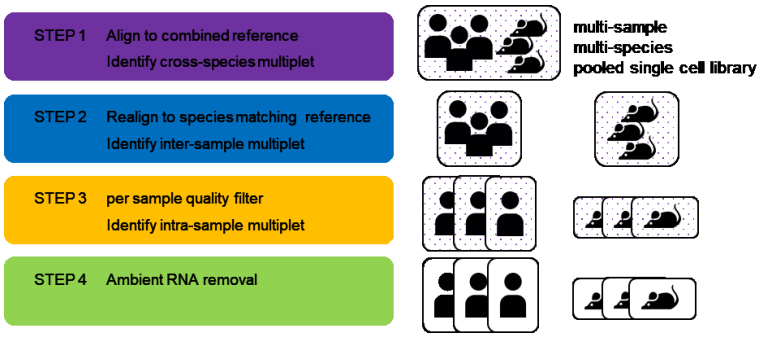



Figure 2. a) pipeline schema

First, we align mixed data to combined reference and identify species of each barcode as well as cross-species doublet. We apply minimum quality filter thresholds during this step as the quality distribution for human and mouse cells may differ. In order to identify cross-species doublet, we used transcriptome wide expression levels to calculate correlation to known single cell data. As misalignment occur in a proportion of genes, the total expression would be more robust to error. True cross-species doublet would bear both expression patterns of human cell and mouse cell. We note that although barcodes with low read counts did not reach correlation values for doublet identification, more than 90% of cells with low correlation value were filtered in future quality filtration step.

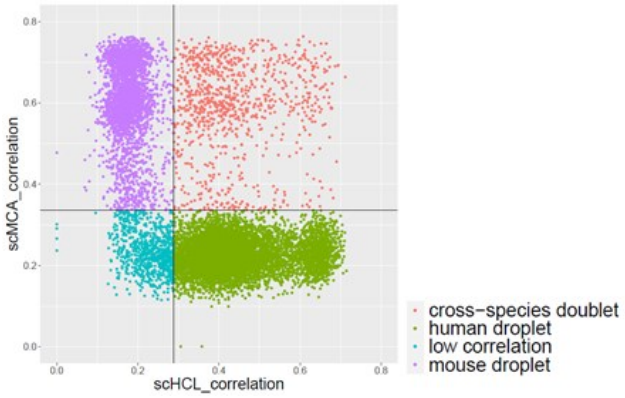
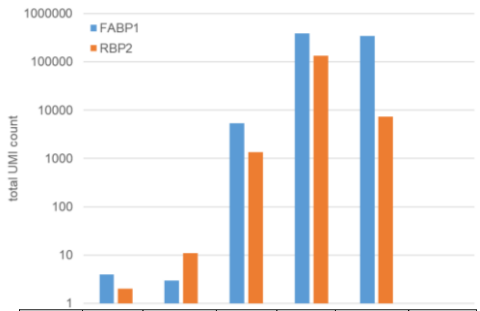



Figure 2. b) Identification of cross-species doublet by expression correlation. Black horizontal and vertical lines are thresholds for doublet identification calculated by normalized median absolute deviation.

After removing cross-species doublets, we realign the data to each species matching reference and select barcodes of the matching species. When multiple identities are pooled, we used variant-based deconvolution method to assign identity for each droplet and detect inter-sample doublets generated from different identities. We apply quality filters for each identity to account for sample level quality variation, then use doublet detection algorithm based on artificial doublet generation to identify intra-sample doublets originating from the same identity.

Although we used species matching reference to remove misalignment error, human singlet cells sequenced from mixed data contained reads which were not expressed in separately sequenced human cell data, namely FABP1 and RBP2. Reads aligning to these genes were found in ambient RNA of the mixed data and also in separately sequenced mouse data aligned to human reference but not in ambient RNA of separately sequenced human data. This suggests that mouse RNA were introduced to human cells through ambient RNA and misaligned to human genes.




| sample | lung tissue | gastric organoid | mixed | mixed | intestinal organoid | ambient RNA removal |
|---|---|---|---|---|---|---|
| library | human | human | mixed | mixed | mouse | mixed |
| species | human | human | human | mouse | mouse | human |
| type | cell | cell | cell | cell | cell | cell |

Figure 2. c) Misaligned reads generated from ambient RNA of different species

Since ambient RNA from the other species misalign similarly in empty droplets and human cells,

methods which remove ambient RNA signals based on expression levels of empty droplets will remove ambient RNA signals generated from both species. We conservatively removed ambient RNA with contamination rate calculated by adding contamination level calculated from each species with an additional 0.05 for strict removal. After ambient RNA removal, all false signals in FABP1 and RBP2 were removed. When comparing the result of our pipeline to the output from CellRanger, our pipeline processed mixed-library data clusters more evenly with separately sequenced data.

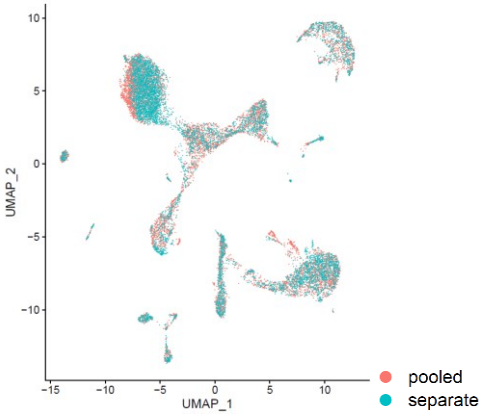



Figure 2 d) Merged UMAP of human cells processed with CellRanger. Cross-species barcodes called by Cellranger and low quality barcodes were filtered.

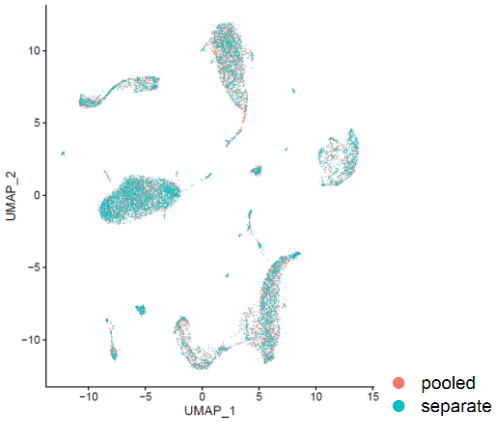



Figure 2 e) Merged UMAP of human cells processed with our pipeline.

We then applied our pipeline to another library containing breast cancer patient derived xenograft tissue and multiple organoids. Since the expression of tumor cells do not resemble tissue originated normal cells, we additionally used copy number clustering method in cross-species doublet identification step. From 23642 initial cells called by CellRanger, we identified 6738 human cells from 6 different identities, and 7580 mouse cells from 2 different identities. Tumor cells (xBC42) had high level of aneuploidy, and the copy number pattern was identical to exome sequencing derived data. Four human identities are lung organoids, three normal organoids (hLO13, hLO18, hLO19) clustered

evenly while one treated with IL13 (hLO11) had distinct cluster. The last human identity was breast organoid (hBO25), which had fewer cells than expected - probably due to slow growth rate. Two mouse identities are microenvironment of xenograft sample (xME42) and mouse pancreatic islet organoid (mPO31). Diverse tumor microenvironment were well captured, two groups of macrophages, endothelial, stromal, muscle cells as well as normal alveolar, luminal, myoepithelial cells were identified.

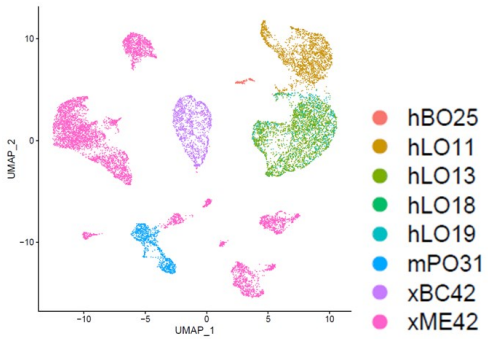



Figure 3 a) Cell identity after REMS pipeline

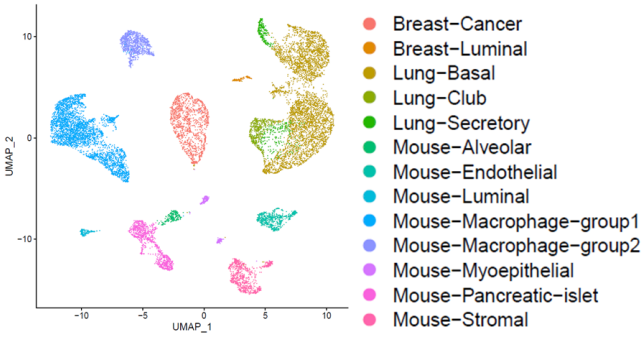



Figure 3 b) Cell type annotation

***Discussion***

When analyzing mixed species single cell data using combined reference, misalignment error greatly affect the result for some genes. The error were not identical across tissues or cell types, and few cells had more than 10% of reads misaligned to the other species. Due to misalignment, cross-species doublet identification via read count or by read percentage[5,21] may lead to inaccurate calls. Also ambient RNA from other species generate strong false positive signals and may disturb downstream analysis.

To cope with misalignment error, we highlight three major consideration points. First, as each droplet contains one cell, the origin of all reads from a droplet should be the same. Identifying species for each droplet while confidently removing cross-species doublet is required. Second, realigning each

droplet to the corresponding reference is required to minimize misalignment error and multiple-mapping events. Third, ambient RNA should be adequately estimated and should be removed.

We developed a pipeline using expression-based species deconvolution which is robust to misalignment errors. Our method can handle pooled library, which can be used to generate cost efficient single cell libraries. Overloading single cell library to target 40,000 droplets, accompanied by adequate doublet removal can reduce the cost of library generation cost to 1/10 per singlet. Our pipeline is not restricted to any certain tool, and one can use other methods as long as they manage misalignment error and adequately remove ambient RNA signals.

Multiple deconvolution methods with or without external treatment are available for multi-sample pooling. CITE-seq[22] uses oligo-tagged antibodies which target commonly expressed cell surface proteins. MULTI-seq[23] uses similar method but with lipid-tagged indices which merges into the membrane of cell or nuclei. Variant based method used natural genetic variation sequenced along with RNA reads to identify each cell[4,18]. But most methods cannot be applied to xenograft samples or co-cultured samples, where mouse and human cells are premixed. Lipid-tagged barcodes are agnostic to species and currently there is no variant based deconvolution tool which handles mixed species data. Cell barcode antibodies which targets human cells or mouse cells separately, or conventional cell sorting technique can be used to separate mixed cells[24]. But both methods need additional process which introduces stress to cells. We made effort to set an unbiased approach with minimal treatment.

During pipeline development, we found some droplets had high mitochondrial gene percentage of the other species, but was not detected as cross-species doublet by multiple methods. The mitochondrial gene percentage were much higher than the amount in ambient RNA, and the possibility of ambient RNA owing to the expression was low. We hypothesize that small cell debris may get mixed into droplet and generate partial cross-species doublets which may be much harder to detect than cross-species doublets. Cell debris from the same species may also form partial doublets and may be mistaken as novel cell types or transient cell state. We handled cells swiftly by an experienced personel to minimize stress and strictly filtered all cell with high mitochondrial percentage. This occurrence would need further evidence with well controlled experiments.

In summary, we point out misalignment error while using combined reference, and present our pipeline which is robust to misalignment to genetrate error corrected gene matrix for downstream analysis. Tools for mixed species single cell data are lacking, and we encourage researchers to optimize our concept with advanced computational method.

***Acknowledgement***

*Funding*

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2020R1A2C3012524) and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2020R1A6A1A03047972)

*Conflict of interests*

The authors declare that they have no competing interests.

1 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).

2 Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381, doi:10.1038/s41586-018-0394-6 (2018).

3 Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **20**, 1349-1360, doi:10.1038/s41556-018-0236-7 (2018).

4 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).

5 Lee, H. W. *et al.* Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. *Genome Med.* **12**, 47, doi:10.1186/s13073-020-00741-6 (2020).

6 Conway, T. *et al.* Xenome--a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, i172-178, doi:10.1093/bioinformatics/bts236 (2012).

7 Kluin, R. J. C. *et al.* XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinformatics* **19**, 366, doi:10.1186/s12859-018-2353-5 (2018).

8 Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347-353, doi:10.1038/s41586-018-0698-6 (2018).

9 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383, doi:10.1038/s41467-018-06318-7 (2018).

10 Stewart, B. J. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365**,

1461-1466, doi:10.1126/science.aat5031 (2019).

11 Kinchen, J. *et al.* Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. *Cell* **175**, 372-386 e317, doi:10.1016/j.cell.2018.08.067 (2018).

12 Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191-W198, doi:10.1093/nar/gkz369 (2019).

13 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

14 Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63, doi:10.1186/s13059-019-1662-y (2019).

15 Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303-309, doi:10.1038/s41586-020-2157-4 (2020).

16 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017, doi:10.1016/j.cell.2018.02.001 (2018).

17 Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.*, doi:10.1038/s41587-020-00795-2 (2021).

18 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).

19 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).

20 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, doi:10.1093/gigascience/giaa151 (2020).

21 Cheloni, S., Hillje, R., Luzi, L., Pelicci, P. G. & Gatti, E. XenoCell: classification of cellular barcodes in single cell experiments from xenograft samples. *BMC Med. Genomics* **14**, 34, doi:10.1186/s12920-021-00872-8 (2021).

22 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).

23 McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nature Methods*, doi:10.1101/387241 (2018).

24 Davis, R. T. *et al.* Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat. Cell Biol.* **22**, 310-320, doi:10.1038/s41556-020-0477-0 (2020).