BMC Bioinformatics

RESEARCH

Open Access

Expression-based species deconvolution and realignment removes misalignment error in multispecies single-cell data

Jaeyong Choi[1,2†], Woochan Lee[1,2†], Jung-Ki Yoon[3], Sun Mi Choi[3,4], Chang-Hoon Lee[3], Hyeong-Gon Moon[5], Sukki Cho[6], Jin-Haeng Chung[7], Han-Kwang Yang[5] and Jong-Il Kim[1,2*]

*Correspondence: jongil@snu.ac.kr
†Jaeyong Choi and Woochan Lee contributed equally to this work.
[1] Department of Biomedical Sciences, Seoul National University College of Medicine, 103, Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea
Full list of author information is available at the end of the article

Abstract

Background: Although single-cell RNA sequencing of xenograft samples has been widely used, no comprehensive bioinformatics pipeline is available for human and mouse mixed single-cell analyses. Considering the numerous homologous genes across the human and mouse genomes, misalignment errors should be evaluated, and a new algorithm is required. We assessed the extents and effects of misalignment errors and exonic multi-mapping events when using human and mouse combined reference data and developed a new bioinformatics pipeline with expression-based species deconvolution to minimize errors. We also evaluated false-positive signals presumed to originate from ambient RNA of the other species and address the importance to computationally remove them.

Result: Error when using combined reference account for an average of 0.78% of total reads, but such reads were concentrated to few genes that were greatly affected. Human and mouse mixed single-cell data, analyzed using our pipeline, clustered well with unmixed data and showed higher k-nearest-neighbor batch effect test and Local Inverse Simpson's Index scores than those derived from Cell Ranger (10 × Genomics). We also applied our pipeline to multispecies multisample single-cell library containing breast cancer xenograft tissue and successfully identified all samples using genomic array and expression. Moreover, diverse cell types in the tumor microenvironment were well captured.

Conclusion: We present our bioinformatics pipeline for mixed human and mouse single-cell data, which can also be applied to pooled libraries to obtain cost-effective single-cell data. We also address misalignment, multi-mapping error, and ambient RNA as a major consideration points when analyzing multispecies single-cell data.

Keywords: Single-cell sequencing, Patient-derived xenograft, Bioinformatics pipeline

Background

Single-cell RNA sequencing is a powerful method used to generate transcriptome with greater resolution [1], identify rare cells [2], and compare cell heterogeneity [3]. Single-cell technology also provides additional functionality such as the ability to discriminate

the expression of human cells from mouse cells in human-mouse xenograft specimens or human-mouse cell coculture 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 data (henceforth called "combined reference"). Cell Ranger (10 × Genomics, Pleasanton, CA, USA) provides a combined reference and flags cellular barcodes where both human and mouse cells are presumably contained in a same droplet (henceforth called "cross-species doublet") using read count distribution. However, similar sequences between each reference may cause human-originated reads to align to the mouse reference [6, 7]. In addition, because many mouse and human genes are homologous, some reads may have the best matching sequence in both references, leading to multi-mapping reads.

We measured the extent of misalignment error and multi-mapping events in single-cell data when using the combined reference to align human and mouse mixed single-cell data. We developed a new realignment pipeline called Realignment- and Expression-based Multispecies deconvolution for Single cell (REMS) to minimize the errors while using minimum prior information about the sequence data. We also expanded and applied REMS to multisample, multispecies pooled single-cell library, enabling us to generate cost-effective single-cell libraries with minimal sequencing batch effects.

Results

Misalignment in the combined reference

To summarize the extent of misalignment in single-cell data, we compared the alignment of the combined reference and the human reference for several human single-cell datasets (Table 1). When aligning human-derived single-cell data to the combined reference, one would expect reads to align to the human part of the combined reference. Most

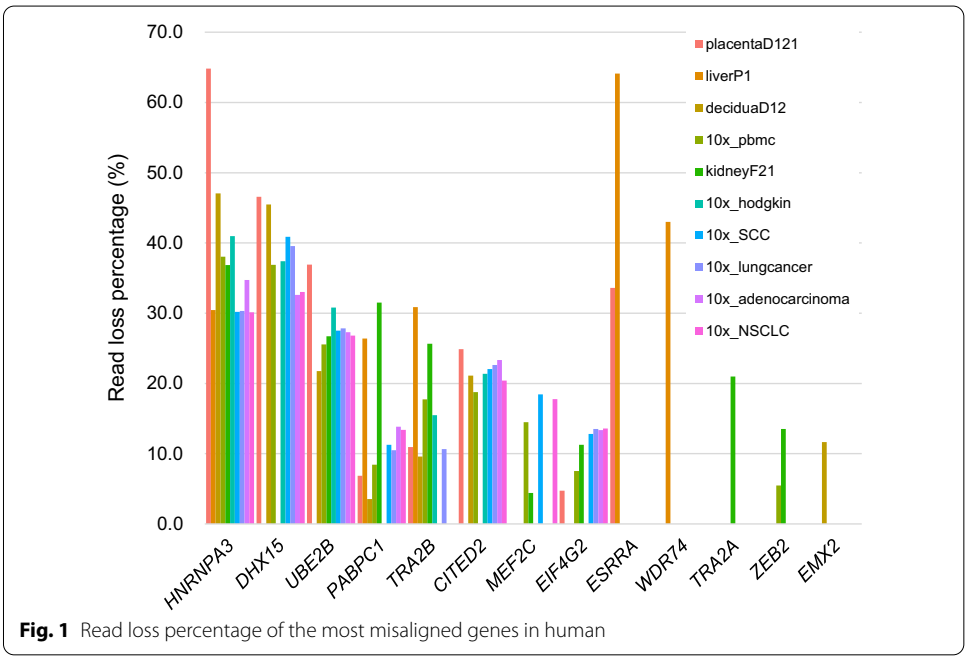

Fig. 1 Read loss percentage of the most misaligned genes in human

Table 1 Misalignment and multi-mapping reads in the combined reference

Data name	Species	Tissue	Cell count	Matching reference		Combined reference			Error in combined reference (%)
				Reads aligned to correct reference	Multi-mapping reads[a]	Reads aligned to correct reference	Reads misaligned to other reference	Multi-mapping reads[a]	
placentaD121	Human	Placenta	2522	35,877,521	1,499,012	35,828,222	11,252	1,665,421	0.50
liverP1	Human	Liver	957	3,179,327	329,018	3,165,784	3243	360,665	1.10
deciduaD12	Human	Decidua	1963	25,147,111	1,186,854	25,122,973	6205	1,333,260	0.61
10x_pbmc	Human	Peripheral blood mononuclear cells	9295	93,171,679	4,730,260	93,027,903	175,980	4,917,645	0.39
kidneyF41	Human	Kidney	3399	21,326,290	1,085,238	21,266,741	77,660	1,150,761	0.67
10x_hodgkin	Human	Hodgkin lymphoma	2502	19,341,273	891,978	19,305,209	22,330	990,901	0.63
10x_SCC	Human	Squamous cell carcinoma	4739	72,819,855	2,515,694	72,583,115	482,564	2,583,374	0.76
10x_lungcancer	Human	Lung cancer	4481	50,316,076	1,927,985	50,165,228	253,394	1,989,662	0.63
10x_adenocarcinoma	Human	Adenocarcinoma	3599	44,820,075	1,562,487	44,678,244	268,975	1,609,641	0.71
10x_NSCLC	Human	Non small cell lung cancer	6640	69,706,720	2,128,698	69,466,298	478,684	2,196,589	0.78
10x_mbrain	Mouse	Brain	8096	103,833,090	18,545,869	103,524,395	38,599	19,909,752	1.35
mcolonHC1	Mouse	Colon	1204	8,810,265	796,909	8,782,774	2440	903,531	1.24

[a] "Multi-mapping reads" are not included in "Reads aligned to correct reference" or "Reads aligned to correct reference"

reads indeed aligned to human genes, but depending on the data type and sequencing throughput, 3000–480,000 reads misaligned to the mouse genes. These misalignments cause expression loss of genes in which the reads should have been mapped. Misalignment errors were more prominent in tumor samples, and errors were also found on aligning mouse-derived single-cell data to the combined reference. Multi-mapping reads, defined as reads which align to more than one exonic locus, also increased 2.7–13.4% in the combined reference compared to the matching reference (Additional file 1). Although multi-mapping read it self is not included in the gene expression, increase in multi-mapping reads also indicates loss of gene expression. All together, when comparing data analyzed using the combined reference and the human reference, 13,000–300,000 fewer reads aligned to human genes in the combined reference than in the human reference.

While all error in combined reference accounted for only 0.4–1.4% of total reads, these reads were concentrated to few genes, leading to strong false signals (Fig. 1, Additional

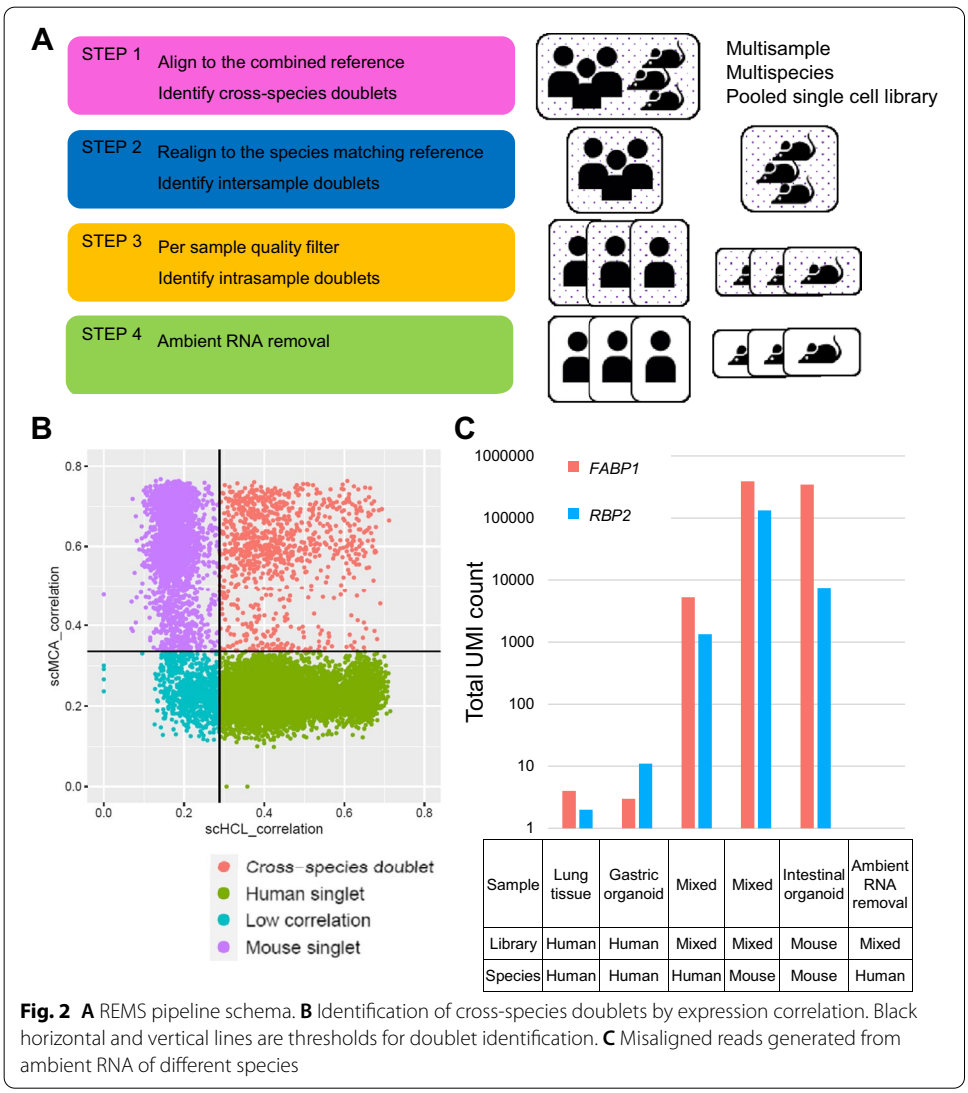

Fig. 2 **A** REMS pipeline schema. **B** Identification of cross-species doublets by expression correlation. Black horizontal and vertical lines are thresholds for doublet identification. **C** Misaligned reads generated from ambient RNA of different species

file 2). For example, 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 human leukocyte antigen genes had high rates of misalignment errors. Interestingly, the genes affected by misalignment and the degree of misalignment were not identical between samples and tissues. Genes with read gain and genes with read loss did not overlap for human and mouse data, suggesting that misalignment was not a simple switch-of-position.

Mixed data analysis pipeline

To assess the misalignment error in human and mouse mixed single-cell data, we generated both mixed and separate libraries of human lung tissue, human gastric organoids, and mouse intestinal organoids. Our mixed data also had misalignment error when we aligned the data to the combined reference (Additional file 3). The error was higher, 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 (Fig. 2A, flowchart available in Additional file 4).

First, we aligned mixed data to the combined reference and identified species and cross-species doublets. Then, we applied a minimum quality filter during the first step, because the quality distribution for human and mouse cells may differ widely. To identify cross-species doublets, we used transcriptome-wide expression levels to calculate the correlation to known single-cell data (Fig. 2B). As misalignment occurs for a proportion of genes, the overall expression pattern would be more robust to error. True cross-species doublets would show expression patterns of both human and mouse cells. We note that although barcodes with low read counts did not reach the threshold for doublet identification, > 90% of all cells with a low correlation value were filtered in future steps.

After removing cross-species doublets, we realigned the data to each species matching reference and selected barcodes of the corresponding species. When multiple identities were pooled, we used a variant-based deconvolution method to assign identity to each droplet and detect intersample doublets generated from different cell identities. After doublet filtration, we applied quality filters for each identity to account for sample level quality variation, and then used a doublet detection algorithm based on artificial doublet generation to identify intrasample doublets originating from the same identity.

Although we used species matching reference to remove misalignment errors, human singlet cells sequenced from mixed data contained reads that were not expressed in separately sequenced human cell data, namely *FABP1* and *RBP2* (Fig. 2C). Reads aligning to these genes were found in ambient RNA of the mixed data and in separately sequenced mouse data aligned to the human reference, but not in ambient RNA of separately sequenced human data. These findings suggested that mouse RNA was introduced into human cell containing droplets in the form of ambient RNA, and it misaligned to human genes.

Since ambient RNA generated reads from the other species misalign similarly in case of empty droplets and human cells, methods that remove ambient RNA signals based on expression levels of empty droplets will remove ambient RNA signals generated from both species. We strictly removed ambient RNA using a contamination rate derived by

Table 2 Comparison of integration metrics between Cell Ranger and REMS pipeline. 0 indicates no integration and 1 indicates full integration

Data type	kBET			LISI		
	Cell ranger	REMS	Difference	Cell ranger	REMS	Difference
mIO04	0.80	0.80	0.00	0.95	0.96	0.00
hGO03	0.11	0.84	0.73	0.91	0.94	0.03
hLT02	0.25	0.36	0.10	0.92	0.95	0.03
T cell (merged)	0.91	0.01	− 0.90	0.91	0.94	0.04
Macrophage	0.62	0.91	0.29	0.92	0.95	0.03
Ciliated	0.32	0.57	0.25	0.93	0.93	0.01
Neutrophil	0.80	0.99	0.19	0.95	0.95	-0.01
Monocyte	0.92	0.98	0.07	0.94	0.96	0.02
Endothelial_1	0.99	0.97	− 0.03	0.93	0.96	0.03
Mast	0.98	0.97	− 0.01	0.92	0.97	0.04
Endothelial_2	0.99	0.97	− 0.02	0.97	0.94	-0.02
Alveolar Type 2	0.94	0.98	0.03	0.94	0.97	0.03

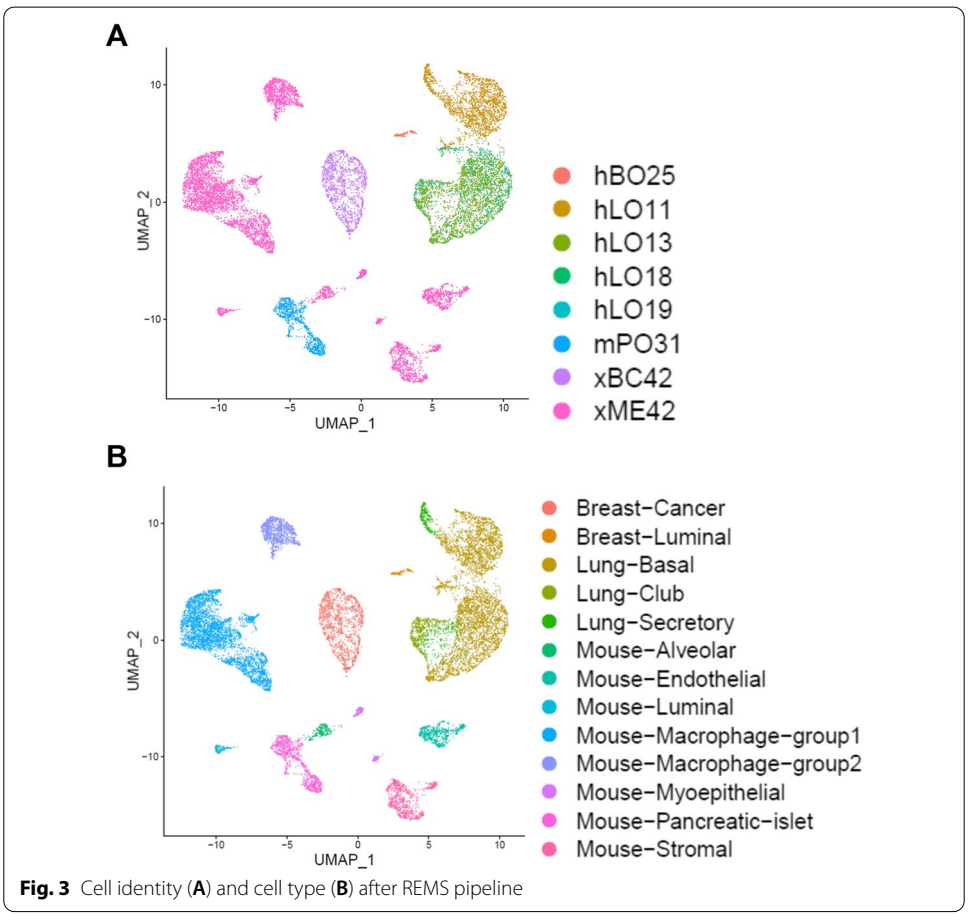

Fig. 3 Cell identity (**A**) and cell type (**B**) after REMS pipeline

adding contamination level calculated from each species with an additional 0.05. After ambient RNA removal, all false signals in *FABP1* and *RBP2* were removed.

To examine the downstream effect of misalignment, we compared the integration level between mixed and unmixed data for the data generated by Cell Ranger and that generated by REMS pipeline (Additional file 5). A total of 98.9% of cells clustered to identical cell types for each dataset, suggesting a small overall effect of misalignment error. However, some samples and cell types had drastic changes in integration metrics. Even after batch effect correction using Harmony, a large difference in the kBET was found for gastric organoid tissue and macrophages, ciliated cells, and neutrophils in lung tissue (Table 2). The LISI had an overall high value for both datasets; REMS generally had higher integration values.

We then applied our pipeline to another library containing breast cancer patient-derived xenograft tissue and multiple organoids. Because the gene expression of tumor cells may not resemble that of normal cells, we additionally used copy number clustering in the cross-species doublet identification step. In depth comparison of result with and without copy number filtering is presented in Additional file 6. From 23,642 initial cells called by Cell Ranger, we identified 6,738 human cells with six different identities and 7,580 mouse cells with two different identities (Fig. 3A, B). Human tumor cells (xBC42) had a high level of aneuploidy, and the copy number pattern was identical to that in exome sequencing-derived data (Additional file 7). Four of the human cell identities were lung organoids: three normal organoids (hLO13, hLO18, and hLO19) clustered evenly while one treated with interleukin 13 (hLO11) had a distinct cluster. The remaining human cell identity was breast organoid (hBO25); fewer cells than expected had this identity. All human identities had best matching genotypes to the corresponding genomic array data. Two mouse identities are microenvironment of xenograft sample (xME42) and mouse pancreatic islet organoid (mPO31). The diverse tumor microenvironment was well captured. Moreover, two groups of macrophages, endothelial, stromal, and muscle cells as well as normal alveolar, luminal, and myoepithelial cells were identified.

Discussion

When analyzing multispecies single-cell data, misalignment errors and multi-mapping events greatly affect the results for some genes. The errors were not identical across tissues or cell types, and some cells had > 10% of reads misaligned to genes of other species. Due to misalignment, cross-species doublet identification via read count or read majority [5, 8] may lead to inaccurate calls. In addition, ambient RNA from other species can generate strong false-positive signals, which may disturb downstream analysis.

To cope with misalignment error, we highlight three major consideration points. First, the origin of reads in a single droplet should be identical. 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 novel pipeline called REMS using expression-based species deconvolution. Our method can handle pooled library and can be used to generate cost efficient single-cell libraries. Overloading a single-cell library to target 40,000 droplets, accompanied by adequate doublet removal can reduce the cost of library generation to as much

as 1/10 per singlet. Our pipeline is not restricted to any specific tool, and one can use other methods as long as they can minimize misalignment error and adequately remove ambient RNA signals. When cancer cells are pooled together, we used copy number profiling as an additional method to remove cross-species doublets. We highly recommend using copy number filtering if one's point of interest lies in mouse cells, since this step adequately removes cancer containing doublets from mouse data. Also, although we filtered all doublets found with DoubletFinder for simplicity, manual inspection of DoubletFinder result may increase the quality of the final data.

Multiple deconvolution methods with or without external treatment are available for multisample pooling. CITE-seq [9] uses oligo-tagged antibodies that target commonly expressed cell surface proteins. MULTI-seq [10] uses a similar method but with lipid-tagged indices that merge into the membrane of cells or nuclei. Variant-based method uses natural genetic variation sequenced in RNA reads to identify each cell [4, 11].

Most methods cannot be applied to xenograft samples or human and mouse cocultured samples, wherein mouse and human cells cannot be separated. Lipid-tagged barcodes are agnostic to species and currently there is no variant-based deconvolution tool that handles multispecies data. Cell barcode antibodies that target human cells or mouse cells separately or a conventional cell sorting technique can be used to separate mixed cells [12]. However, both methods need additional processing, which induces stress in the cells. We attempted to establish an unbiased approach using minimal treatment.

Currently there is no technology to tag mouse or human RNA on a molecular level. Thus we could not quantify misalignment error in mixed-species single cell data, as we cannot verify the true expression value in the existence of ambient RNA from the other species. We presented extreme cases where expression was lost or non-existing expression was found; but for other genes, we could not quantify the degree of misalignment error.

During the development of REMS, we found that some droplets had high mitochondrial gene percentage of the other species, but they were not detected as cross-species doublets using multiple detection methods. The mitochondrial gene percentage in these droplets was much higher than that in ambient RNA, and the possibility of ambient RNA owing to its expression was low. We hypothesize that small cell debris may get mixed with droplets and generate partial cross-species doublets that 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 states. In this study, cells were handled swiftly by experienced personnel to minimize stress and all cells with high mitochondrial percentage were stringently filtered.

Conclusion

In summary, we identified misalignment error and multi-mapping event while using the combined reference and present our realignment pipeline, which is robust for misalignment to generate error-corrected gene matrix for downstream analysis. Tools for multispecies single-cell data are lacking, and we encourage researchers to optimize our pipeline strategy using advanced computational methods.

Methods

Tissue preparation

Normal human tissues were obtained from early-stage lung, breast, and stomach cancer surgery. Pathologists confirmed the absence of cancer cells. Three samples of human lung tissues were obtained from tissue donations of three patients with idiopathic pulmonary fibrosis who underwent lung transplantation. A bronchoscopic brushing sample was collected from healthy bronchus obtained from a patient who received bronchoscopy. Furthermore, a breast cancer patient-derived xenograft mouse was sacrificed 6 weeks after tumor injection.

Fresh human tissues were treated using the Multi Tissue Dissociation Kit 1 (Miltenyi Biotec, Bergisch Gladbach, Germany) along with the gentleMACS Octo Dissociator (Miltenyi Biotec) and 37C_Multi_A_01 program. The brushing sample was centrifuged at 400*g* for 5 min at 4 °C and the pellet was resuspended in Advanced DMEM/F12 (Gibco, Dublin, Ireland) with Glutamax (Gibco), HEPES (Gibco), Penicillin–Streptomycin (Gibco) (ADF^{+++}), and 1 mg/mL collagenase IV (Sigma, St. Louis, MO, USA), followed by digestion for 15 min at 37 °C.

Patient derived xenograft model was established with female immunodeficient CIEA NOG mouse (NOD.Cg-Prkdc^{scid}IL2γg$^{tm1\ Sug}$/JicKoat, KOATECH) and human tumor. After euthanize mouse with CO2 chamber, Resected xenograft tumor tissue was dissociated into a single cell suspension using the Tumor Dissociation Kit 1, human (Miltenyi Biotec) along with the 37C_h_TDK_3 program. Mouse normal islet and intestine tissues were obtained from a 12-week-old female C57BL/6 mouse. Mouse islet cells were purified using Ficoll-Paque (Cytiva, Marlborough, MA, USA) and digested using TrypLE™ (Gibco) for 15 min at 37 °C. Mouse intestinal crypt cells were isolated using the Gentle Cell Dissociation Reagent (STEMCELL Technologies, Vancouver, Canada) for 20 min at room temperature followed by squeezing using a microscope cover glass.

Each digested tissue (without intestinal crypt cells) suspension was strained using a 70-μm filter. The suspension was then centrifuged at 400*g* for 5 min at 4 °C with 10 mL of ADF^{+++}. The pellet was resuspended in 1 mL of 1 × Red Blood Cell Lysis Solution (Miltenyi Biotec) for 10 min at room temperature for lysis of erythrocytes, followed by the addition of ADF^{+++} and centrifugation at 400*g* for 5 min. The pellet was then resuspended in an appropriate volume of ADF^{+++} and kept on ice for organoid culture or library generation.

Organoid culture

After digestion and erythrocyte lysis, the cell pellets were seeded in Corning® Matrigel® Growth Factor Reduced Basement Membrane Matrix (Corning, Corning, NY, USA), solidified for 15 min at 37 °C, and cultured on each organ-specific media for human lung [13], human breast [14], human gastric [15], mouse islet [16], and mouse intestine [17] cells. The media was changed every 2–3 days and the tissues were subcultured every 2–3 weeks depending on the density and growth rates of organoids.

Single-cell RNA sequencing

We performed single-cell RNA sequencing according to the standard 10 × Genomics 3′ V3.1 chemistry protocol (10 × Genomics). Before loading cells on the G chip, cells were

counted to determine the cell concentration. The number of loading cells per library was calculated via the multiplexing cost calculator (satijalab.org/costpercell/). The libraries were sequenced using the NovaSeq 6000 (Illumina, San Diego, CA, USA) platform S4 with paired-end 100 bp.

Assessment of misalignment

We aligned the data to species matching reference data (refdata-gex-GRCh38-2020-A or refdata-gex-mm10-2020-A) using Cell Ranger 4.0.0 (10 × Genomics). Seurat v3.2.3 [18] was used to filter barcodes by low total read count, low expressed gene count, and high mitochondrial gene percentage. We extracted these cellular barcodes from the data aligned to the combined reference (refdata-gex-GRCh38-and-mm10-2020-A) and compared the alignment result. We defined misaligned read as the following: (1) read aligned to a gene of the other species or (2) read difference between the combined reference and the species matching reference. We defined "misaligned gene" as the following: (1) average read count higher than 0.1 per cell and (2) > 10% difference in the read count between the species matching reference and the combined reference. Pathway analysis for misaligned genes was performed using the web server g:Profiler [19].

We defined multi-mapping reads as reads which align to more than one exonic loci, since reads mapped to intronic and intergenic region does not contribute to expression count matrix. In detail, we excluded reads of the following criteria; (1) reads mapped to multiple non-exonic loci; (2) read mapped to multiple loci with only one exonic locus, since Cell Ranger adjusts the mapping quality and aligns the read to the exonic locus; (3) reads mapped to multiple loci, but primarily to the other species. We calculated the error in combined reference as the sum of (1) reads misaligned to other species and (2) the increment of multi-mapping reads; compared to total reads in the matching reference. The total count of align reads may differ for each reference due to (1) difference in multi-mapping reads; (2) reads mapped to non-exonic loci in one species may map to an exonic region of the other species, which increases the count of the other species.

REMS pipeline step 1: Alignment and prefilter

We aligned human and mouse mixed single-cell data to the combined reference using Cell Ranger 4.0.0. Seurat v3.2.3 [18] was used to filter cells, normalize expression, cluster, and visualize data. The EmptyDrops method [20] was used to identify droplets that had an expression pattern similar to that of ambient RNA. We defined ambient RNA as "RNA molecules present in cell containing droplets, which does not originate from the cell". The expression of ambient RNA was calculated with barcodes which had less than 100 reads. We selected barcodes with expressed gene count ≥ 200 in the human reference and/or mouse reference. We removed barcodes if their percentage of mitochondrial genes was twice or higher than the percentage of mitochondrial genes in ambient RNA; this was performed separately for human and mouse genes.

REMS pipeline step 2: Expression-based cross-species doublet identification

For data containing nontumor cells, we used the single cell Human Cell Landscape (scHCL) [21] and single cell Mouse Cell Atlas (scMCA) [22] correlation values to estimate cross-species doublets. The cutoff was set to 99% confidence interval of median

normalized correlation value for human and mouse cells. Barcodes with both high correlations to human cells and mouse cells were classified as cross-species doublets. Other barcodes were classified as human cells or mouse cells according to read majority and correlation value. We removed barcodes with discrepancies. When expression for each species could be clearly separated by clustering, minor species in each cluster were removed.

For data containing tumor cells, we used copy number cluster result from CopyKAT [23] in addition to correlation values to identify tumor cells with low correlation value. We selected CopyKAT since it is currently the only program which requires no additional information other than gene expression matrix. We defined aneuploid barcodes as tumor cells. Barcodes with both valid copy number result and mouse read majority were additionally classified as cross-species doublets.

REMS pipeline step 3: Species matching reference realignment

After species identification, we realigned data to the human reference or mouse reference, and selected species matching barcodes. When multiple samples were pooled, we used Souporcell [11] to identify intersample doublets. We selected singlets identified using Souporcell, and when the 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 a 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 [24] was used to identify intrasample doublets.

REMS pipeline step 4: Ambient RNA removal

SoupX [25] was used to estimate the ambient RNA expression from barcodes with ≤ 100 reads. We calculated the contamination rate in human and mouse cells separately. We additionally added 0.05 to the contamination rate to stringently remove ambient RNA signals.

Result comparison

Harmony [26] was used to remove batch effect from mixed and separately sequenced libraries. Cell type was identified by the consensus of scHCL correlation for each cluster. We used the k-nearest-neighbor batch effect test (kBET) [27] with top 20 Harmony components and Local Inverse Simpson's Index (LISI) [26] with Uniform Manifold Approximation and Projection (UMAP) coordinates calculated from Harmony components. We calculated integration metrics for each sample and celltypes with ≥ 100 cells. Each metric was normalized to a 0–1 scale, 0 indicating no integration and 1 indicating full integration.

Genomic array

DNA was extracted from tissue or cultured organoids using DNeasy Blood & Tissue Kit (Qiagen, Germantown, MD, USA). Axiom Korea Biobank Array (ThermoFisher

Scientific, Waltham, MA, USA) [28] was used to genotype samples. Quality control and filtering was performed following the Korea Biobank Array protocol.

Identification of pooled sample

Souporcell generates estimated genotypes for each sample which it deconvoluted. PLINK v1.0.9 identity-by-descent [29] was used along with linkage disequilibrium variants pruned genomic array data as true genotype to confirm single-cell deconvoluted sample identity.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04676-0.

Additional file 1. The percentage of Cell Ranger adjusted reads. When a read is mapped to multiple-loci, with only one exonic locus, Cell Ranger adjusts the mapping quality and counts the read toward the mapped gene. The percentage Cell Ranger adjusted in the other species' reference is much higher than the percentage in the matching species.

Additional file 2. Read loss percentage of the most misaligned genes in mouse.

Additional file 3. Misalignment and multi-mapping reads in combined reference from mixed and separately sequenced data.

Additional file 4. Flowchart of REMS pipeline.

Additional file 5. MAP of human lung cells processed using Cell Ranger annotated by (A) library and (B) cell type. UMAP of human lung cells processed using REMS pipeline annotated by (C) library and (D) cell type.

Additional file 6. Comparison of "REMS with copy number filter" and "REMS without copy number filter".

Additional file 7. Copy numbers from single cell and exome data.

Acknowledgements

We thank the Genomic Medicine Institute Research Service Center for guidance and use of the research computing infrastructure.

Author contributions

Conceptualization: JC, WL; Data Curation: JC, WL, JKY; Formal Analysis: JC, Funding Acquisition: JIK; Investigation JC, WL, JKY; Methodology: JC, WL, JKY; Project Administration: JIK; Resources SMC, CHL, HGM, SC, JHC, HKY, JIK; Software: JC; Supervision: JIK, Validation: JC, WL, JKY; Visualization: JC, Writing: JC. All authors read and approved the final manuscript.

Funding

This work was supported by a National Research Foundation of Korea grant funded by the Korean government (2020R1A2C3012524) and by a Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2020R1A6A1A03047972). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Human decidua (D12) [30], placenta (D121), liver (P1) [31], kidney (F41) [32], and mouse colon (HC1) [33] single-cell data were downloaded from the Human Cell Atlas Data Portal (data.humancellatlas.org). Human blood, Hodgkin lymphoma, pooled cancer, and mouse brain data were downloaded from 10 × Genomics single-cell gene expression datasets (support.10xgenomics.com/single-cell-gene-expression). Mixed and separately sequenced single-cell data are available at Sequence Read Archive under BioProject number PRJNA719675. Xenograft and lung organoid single-cell data are not publicly available due to institutional review board restrictions and are available from the corresponding author on reasonable request. All software used in REMS pipeline is available for public use at; Cell Ranger: support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest, scHCL: github.com/ggjlab/scHCL; scMCA github.com/ggjlab/scMCA, CopyKAT: github.com/navinlabcode/copykat, Souporcell: github.com/wheaton5/souporcell, DoubletFinder: github.com/chris-mcginnis-ucsf/DoubletFinder, SoupX: github.com/constantAmateur/SoupX. Additional information regarding REMS bioinformatics pipeline can be found at the following site [34].

Declarations

Ethics approval and consent to participate

All studies were conducted in accordance with the Declaration of Helsinki and the Guidelines for Good Clinical Practice and performed after receiving written informed consent obtained from each participant and/or their legal representative, as appropriate. The institutional review board of the Seoul National University Hospital granted permission to perform this study (2008-065-1148, 1909-110-1066, 1805-151-948, and 1402-054-555). All animal studies were carried out in accordance with ARRIVE guidelines and other relevant guidelines/regulations. Mouse normal islet and intestine tissues were obtained from a 12-week-old female C57BL/6 mouse. Patient derived xenograft model was established with female

immunodeficient CIEA NOG mouse. The Institute of Laboratory Animal Resources Seoul National University approve all animal experiment in this study (19-0096-S1A1(2), SNU-190401-5-7, SNU-200707-1-1).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biomedical Sciences, Seoul National University College of Medicine, 103, Daehak-ro, Jongno-gu, Seoul 03080, 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 Hospital, Seoul, Republic of Korea. [4]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea. [5]Department of Surgery, Seoul National University College of Medicine, Seoul, Republic of Korea. [6]Department of Thoracic and Cardiovascular Surgery, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [7]Department of Pathology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea.

Received: 11 July 2021 Accepted: 28 March 2022
Published online: 02 May 2022

References

1. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
2. Plasschaert LW, Zilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature. 2018;560(7718):377–81.
3. Lawson DA, Kessenbrock K, Davis RT, Pervolarakis N, Werb Z. Tumour heterogeneity and metastasis at single-cell resolution. Nat Cell Biol. 2018;20(12):1349–60.
4. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36(1):89–94.
5. Lee HW, Chung W, Lee HO, Jeong DE, Jo A, Lim JE, 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. 2020;12(1):47.
6. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, et al. Xenome–a tool for classifying reads from xenograft samples. Bioinformatics. 2012;28(12):i172–8.
7. Kluin RJC, Kemper K, Kuilman T, de Ruiter JR, Iyer V, Forment JV, et al. XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. BMC Bioinform. 2018;19(1):366.
8. Cheloni S, Hillje R, Luzi L, Pelicci PG, Gatti E. XenoCell: classification of cellular barcodes in single cell experiments from xenograft samples. BMC Med Genom. 2021;14(1):34.
9. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19(1):224.
10. McGinnis CS, Patterson DM, Winkler J, Hein MY, Srivastava V, Conrad DN, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods. 2018;16:619–26.
11. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat Methods. 2020;17(6):615–20.
12. Davis RT, Blake K, Ma D, Gabra MBI, Hernandez GA, Phung AT, et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. Nat Cell Biol. 2020;22(3):310–20.
13. Sachs N, Papaspyropoulos A, Zomer-van Ommen DD, Heo I, Bottinger L, Klay D, et al. Long-term expanding human airway organoids for disease modeling. EMBO J. 2019;38(4):e100300.
14. Rosenbluth JM, Schackmann RCJ, Gray GK, Selfors LM, Li CM, Boedicker M, et al. Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. Nat Commun. 2020;11(1):1711.
15. Bartfeld S, Bayram T, van de Wetering M, Huch M, Begthel H, Kujala P, et al. In vitro expansion of human gastric epithelial stem cells and their responses to bacterial infection. Gastroenterology. 2015;148(1):126-U554.
16. Wang D, Wang J, Bai L, Pan H, Feng H, Clevers H, et al. Long-term expansion of pancreatic islet organoids from resident procr(+) progenitors. Cell. 2020;180(6):1198-211.e19.
17. Sato T, Vries RG, Snippert HJ, van de Wetering M, Barker N, Stange DE, et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature. 2009;459(7244):262-U147.
18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-902.e21.
19. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
20. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas J, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20(1):63.
21. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581(7808):303–9.
22. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018;172(5):1091-107.e17.

23. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608.
24. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8(4):329-37.e4.
25. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9(12):giaa151.
26. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96.
27. Buttner M, Miao ZC, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
28. Moon S, Kim YJ, Han S, Hwang MY, Shin DM, Park MY, et al. The Korea Biobank array: design and identification of coding variants associated with blood biochemical traits. Sci Rep. 2019;9(1):1382.
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
30. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature. 2018;563(7731):347–53.
31. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018;9(1):4383.
32. Stewart BJ, Ferdinand JR, Young MD, Mitchell TJ, Loudon KW, Riding AM, et al. Spatiotemporal immune zonation of the human kidney. Science. 2019;365(6460):1461–6.
33. Kinchen J, Chen HH, Parikh K, Antanaviciute A, Jagielowicz M, Fawkner-Corbett D, et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. Cell. 2018;175(2):372-86.e17.
34. REMS pipeline. https://snumrc.snu.ac.kr/gmi/en/community/gallery?md=v&bbsidx=121. Accessed 13 May 2021.

Publisher's Note