**Title page**

# CellClear: Enhancing Single-cell RNA Data Quality via Biologically-Informed Ambient RNA Correction

Wanxiang Huang[1, ‡] | LongFei Hu[1, ‡] | Zhengzong Qian[1] | Jue Fan[1]

[1]Singleron Biotechnologies, Nanjing, China

[‡]The first two authors contributed equally to this work.

**Abstract**

Ambient RNA, caused by the pool of mRNA molecules released from lysed or dead cells in samples, poses persistent challenges to single-cell RNA sequencing data analysis, and leads to potential problems such as inaccurate classification of cell types and downstream functional analyses. Here we propose CellClear to estimate and remove the ambient RNA expression programs in single-cell RNA sequencing data. CellClear demonstrates significant increase in the accuracy of ambient gene expression correction comparing with state-of-art methods, without distorting native information of individual cell types.

**Keywords**

Ambient RNA, scRNA-seq, Non-Negative Matrix Factorization, Contamination Removal

## Background

Single-cell RNA sequencing (scRNA-seq) is a transformative advancement that unveils cellular heterogeneity within complex tissues at single-cell resolution. It facilitates the identification of rare cell types, and allows for the characterization of dynamic cellular states [1,2]. Despite its potentials, scRNA-seq technology can suffer from technical noises, such as low RNA capture efficiency and the presence of ambient RNA [3,4]. Ambient RNA originating from ruptured, dead, or dying cells, as well as cell-free RNA including exome RNA, results in non-zero counts in cell-free barcodes and non-specific gene expressions in cell-associated barcodes [5].

Currently, various tools have been developed for ambient RNA correction. DecontX assumes cells as mixtures of native and contamination-derived counts and iteratively refines their relative proportions [4]. SoupX identifies cluster-specific markers to estimate the fraction of Unique Molecular Identifiers (UMIs) originating from the background and use it to represent the ambient RNA fraction. It is recommended to complement this automated estimation with a set of contamination-related markers for manual estimation [6]. CellBender, a deep generative model based on a neural network, performs well in most datasets [5], but demands significantly high computational costs [7,8]. However, with the increasing trend of analyzing large scale scRNA-seq datasets, there is an urgent need for a method that can not only rapidly and efficiently assess the necessity for decontamination during the quality control stage without prior knowledge, but also accurately correct ambient expression across various data situations such as complex contamination sources.

Here, we introduce CellClear, which can accurately identify and correct ambient genes while preserving the biological features of the data. The corrected data shows a significant improvement of sensitivity during the downstream function analyses. To avoid unnecessary

corrections, CellClear also provides an ambient expression level as a QC metric to guide researchers in deciding whether to apply the correction. Comparing to existing methods, CellClear shows superior performance in ambient RNA correction, making it an efficient tool to be included in standard scRNA data analysis pipelines.

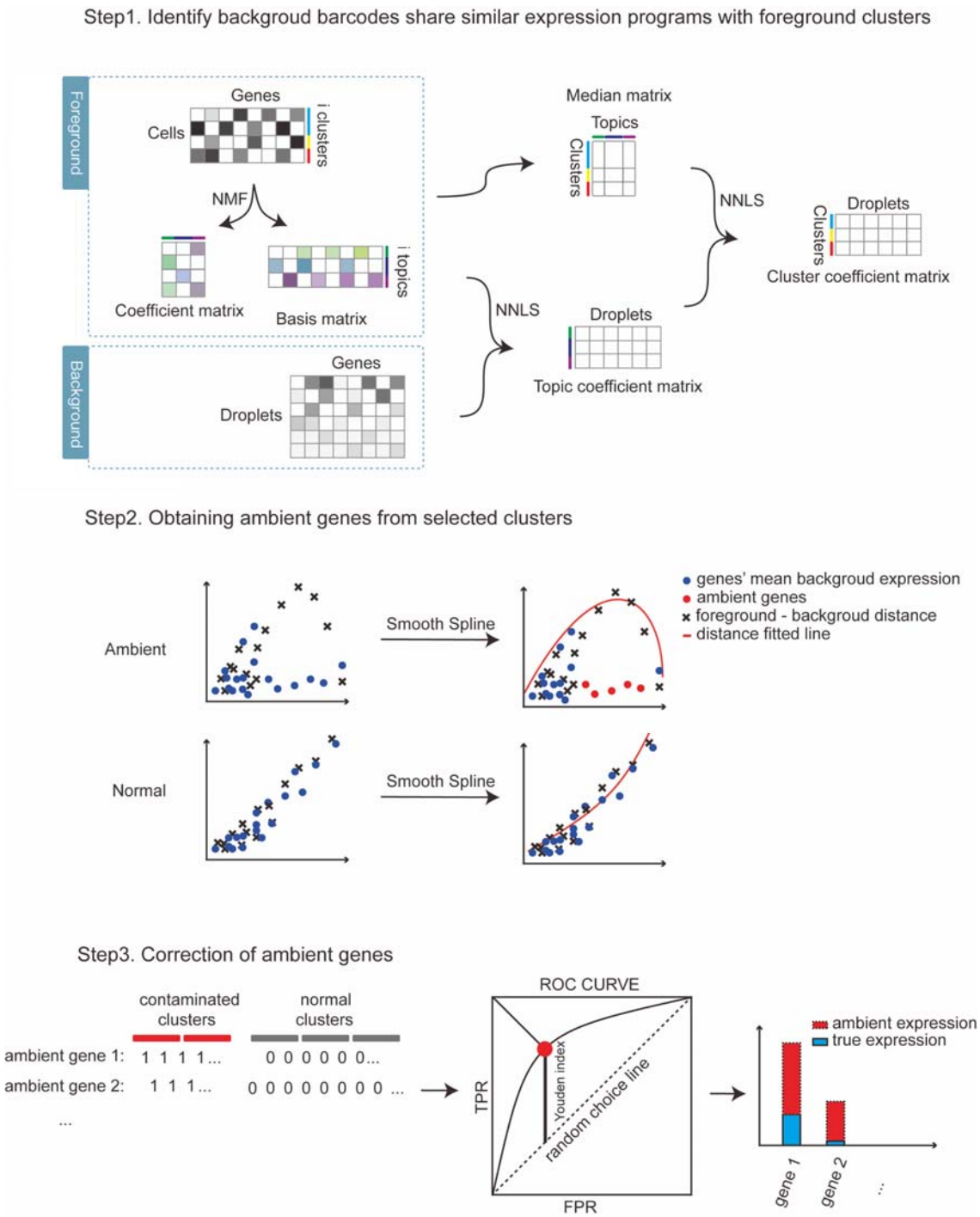


Figure 1. Overview of the CellClear method.
Step1. Non-Negative Matrix Factorization (NMF) and two-step Non-Negative Least Squares (NNLS) optimisation are applied to foreground and background matrices. Step2. Genes with abnormal foreground-background distances are taken as ambient genes. Step3. Area Under the Receiver

Operating Characteristic Curve (AUC-ROC) and Youden index are used to determine the threshold for each ambient gene.

## Results and Discussion

To showcase CellClear's performance, we conducted evaluations using three scRNA-seq replicates from mouse kidney samples [8]. These samples, known for elevated ambient expression levels in Proximal tubule cells (PTs) and specific ambient genes like Slc34a1, were utilized for benchmarking CellClear against previously published tools.

To assess whether the ambient correction process accurately corrected ambient genes, the corrected data from four tools were compared (Figure 2A, Additional file 1: Figure S1A). Both CellClear and CellBender consistently and effectively reduced the expressions of Slc34a1 in non-PT cells across all three samples (Figure 2B, Additional file 1: Figure S1B). DecontX showed variable performance, effectively reducing Slc34a1 expressions in two of the samples but performing poorly in the other, whereas SoupX appeared to fail to accurately remove non-specific expressions of Slc34a1 in all replicates.

In the context of ambient gene correction, it is crucial not only to capture ambient genes but also to ensure that the other genes remain unaffected, thereby preserving the biological feature of the data. Moreover, ambient expressions typically are from not just one gene but multiple genes that share similar expression patterns. The ambient expression pattern in this dataset, mainly composed of the top 10 PT markers, was expected to be corrected by four tools, resulting in increased sensitivity of detecting these PT markers. After correction by CellClear, CellBender, and DecontX, the log2 fold changes (log2FC) of the top 10 PTs markers in PT cells versus all other cells were significant increased comparing to the uncorrected data. Meanwhile, both CellClear and CellBender effectively reduced the expression percentage of these markers in non-PTs cells (Figure 2C). For non-ambient genes, the expression levels and percentages of intercalated cells of collecting duct and Endothelial markers remained largely unchanged across all four tools (Additional file 1: Figure S1C).

To evaluate the precision of these tools in detecting ambient RNA expression, the estimated ambient expression fraction per cell provided by each tool were compared, using genotype-based ambient expression estimates in the original publication as the ground truth. As illustrated in Figure 2D, CellClear consistently outperformed the others across the three replicates, highlighting the close alignment of CellClear's ambient expression fractions in each cell with the genotype-based estimates.

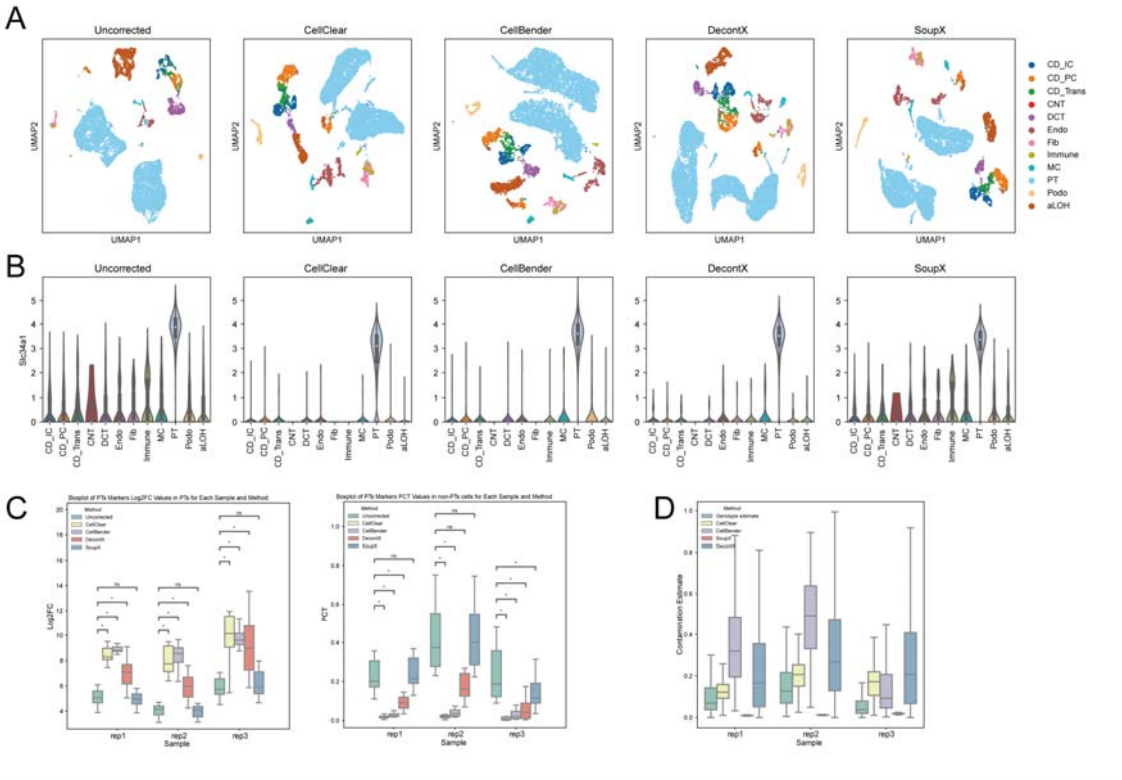


Figure 2. Performance Evaluation of Ambient RNA Correction Tools on Mouse Kidney Samples
(A) UMAP scatter plots of mouse kidney replicate 1 before and after ambient expression correction, coloured by cell type labels (PT, proximal tubule cells; CD_IC, intercalated cells of collecting duct; CD_PC, principal cells of collecting duct; CD_Trans, transitional cells of collecting duct; CNT, connecting tubule cells; DCT, distal convoluted tubule cells; Endo, endothelial cells; Fib, fibroblasts; aLOH, ascending loop of Henle cells; dLOH, descending loop of Henle cells; MC, mesangial cells; Podo, podocytes). (B) Violin plots showing the expression levels of the PTs marker Slc34a1 before and after correction by four tools. (C) Comparison of log2FC and percentage of expression (PCT) before and after correction by four tools in differential expression analysis of top 10 PT markers. (D) Estimated ambient expression levels per cell based on genetic variants and four ambient correction tools.

We next sought to understand how corrected data affect downstream analysis in a public human PBMC dataset, also used by CellBender to demonstrate its performance [5]. The cell type of each cluster in corrected and uncorrected data was identified (Figure 3A) based on the markers provided in the article [5]. CellClear showed an ambient expression level of 0.22 and the top 20 ambient genes included Neutrophil/Monocytes-related genes (S100A8/A9/LYZ), and HLA genes (Additional file 2: Table 1). The widespread expression of Neutrophil/Monocytes-related genes in all cell types of uncorrected data , their significantly reduced expression in CellClear-corrected data (and ) and reduced but to a less degree in CellBender-corrected data (90.5% expressed in non-Monocytes cells versus 0.2% and 7.7%) demonstrate CellClear's accuracy in identifying and correcting ambient genes (Figure 3B, Additional file 1: Figure S2A, S2B).

The corrected data from two tools were then used to assess the impact on downstream analyses. Differential gene expression and Gene Ontology enrichment analyses [9] were performed to verify if corrected data could offer a more precise functional interpretation than uncorrected data. Pathway enrichment analyses were compared using the top 10 shared

pathways enriched by up-regulated genes from each cell type in uncorrected and corrected data. The analyses showed that in the CellClear-corrected data there was a significant increase in the level of enrichment, particularly for T cell subtypes that were significantly affected by the ambient expression of Neutrophil/Monocyte related genes (Figure 3C). This demonstrates that correction may enhance the sensitivity of pathway analysis for each cell type

Cell-cell interactions could also be significantly affected by ambient expression. After correction by CellClear, CellphoneDB [10] analysis revealed a notable decrease in the number of interactions between cell types, with the total number of significant interacting pairs decreasing from 1815 to 1233 (Figure 3D). The data corrected by CellBender showed a similar pattern but to a less degree, decreasing from 1815 to 1564. This result suggested that inflated background noise due to ambient expression could potentially cause some of the interactions to be incorrectly detected by cell communication tools such as CellphoneDB. For example, previous study has shown the interaction pair CD55-ADGRE5 between Monocytes and T cells plays a crucial role in T cell activation [11]. This particular interaction was falsely identified as a pair between B cells and T cells due to Monocytes-related contamination in the uncorrected data, while CellClear corrected data reduced this false positive.

In summary, CellClear outperformed other tools in identifying and correcting ambient genes efficiently, preserving the original biological features and improving the accuracy of downstream analysis.

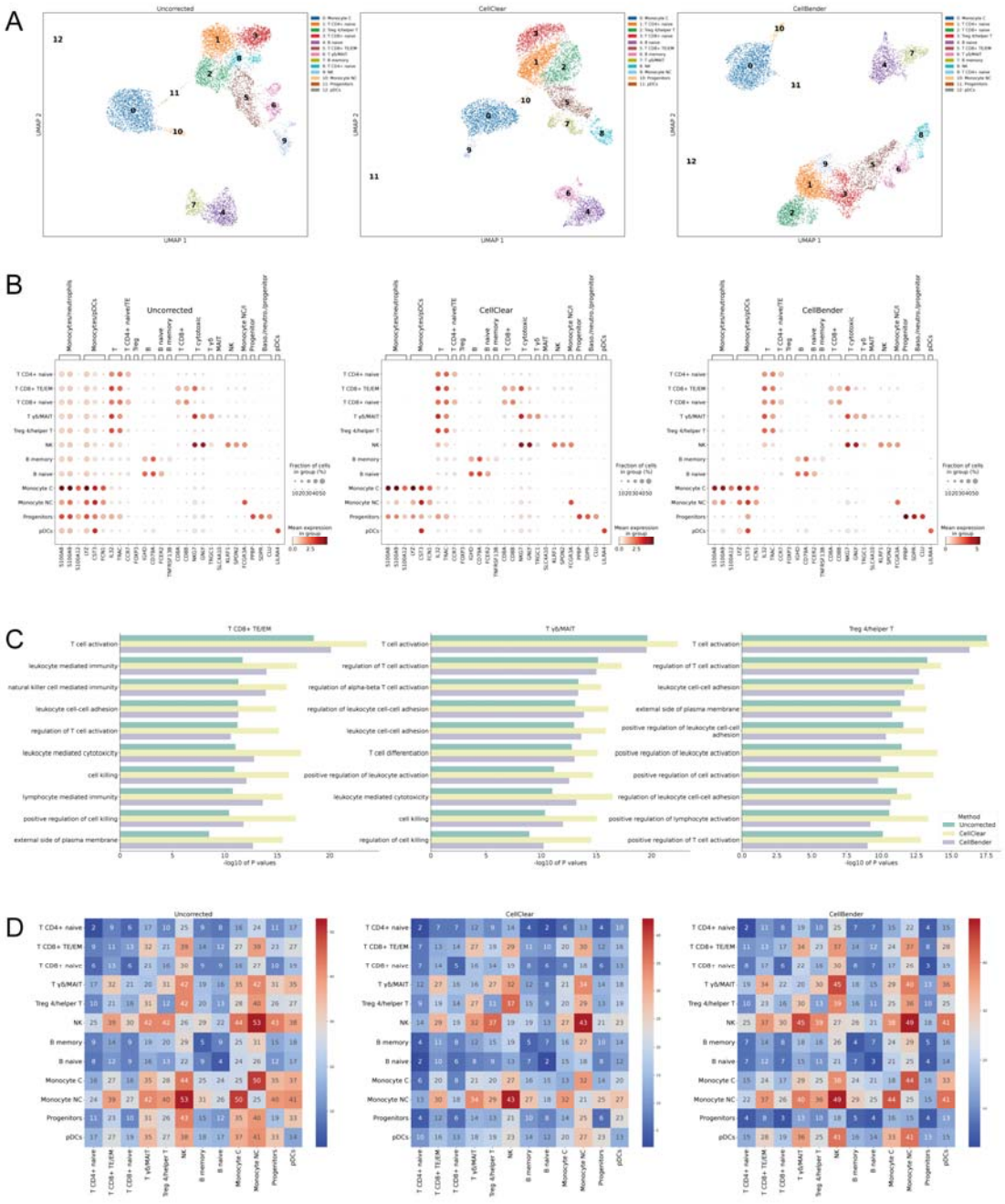


Figure 3. Comparison of the Potential Impacts of Ambient RNA Correction Tools on Downstream Analyses

(A) UMAP scatter plots of a public human PBMC sample before and after ambient expression correction, coloured by cluster labels (Baso., basophil; neutro., neutrophil; NK, natural killer cell; Treg, regulatory T cell; MAIT, mucosal-associated invariant T cell; monocyte C, classical monocyte; monocyte NC, non-classical monocyte). (B) Dot plots showing the expression levels of markers before and after correction. (C) Comparison of -log10 (p-values) for shared top 10 pathway enrichment across each cell type before and after correction, using differentially expressed genes with

log2FC ⩾ 1. (D) Heatmap showing the number of interaction pairs before and after correction. Color bar from blue to red showed the number of interacting pairs between the two cell types from low to high.

## Conclusions

CellClear distinguishes itself from existing tools like DecontX, SoupX and CellBender by its superior capability to correct ambient genes while preserving biological features. The top ambient genes and their contamination levels provided by CellClear enable researchers to quickly assess the data during the quality control stage, identify the primary contamination sources, and correct them when necessary. Comparing to the discovery of data quality issues related to ambient RNA after the data integration step, this method avoids the tedious work to iterate between QC and data integration. Furthermore, the corrected data demonstrated cell types with more specific differentially expressed genes and significantly enriched key pathways, revealing its capability for enhancing downstream analysis. All these evidences substantiate the assertion that CellClear is a reliable tool for restoring true expression programs by selectively removing ambient noises.

## Methods

### Identify background barcodes that exhibit similar expression programs to the foreground clusters.

The CellClear method employs clustering and Non-Negative Matrix Factorization (NMF) to derive cluster-relevant expression programs from foreground cell matrix, which is the cell associated matrix identified by primary analysis pipelines. The W and H matrices of NMF are initially set using the cellular identities of individual cells and the log fold changes of differentially expressed genes identified by Scanpy [12]. A clustering resolution range of 0.8 to 1.2 is recommended to obtain more cluster-relevant expression programs. This approach allows us to seed the model with prior information (cluster-specific programs), reduce the risk of overfitting during training and guide the model towards biologically relevant results [13]**Error! Reference source not found.**.

Next, the unit variance scaled foreground counts matrix $V_{km}$, which has m cells and k features, is infinitely approximated with $\widetilde{V_{km}}$:

$$V_{km} \approx \widetilde{V_{km}} = W_{ki} \cdot H_{im}, (W_{ki} \geq 0, H_{im} \geq 0)$$

$W_{ki}$ is the basis matrix, with each row representing a basis vector or an inherent component of the original data. $H_{im}$ is the coefficient matrix, and each column of matrix $H_{im}$ indicates the coefficients corresponding to the basis vectors in $W_{ki}$. The index i represents the number of cluster-relevant expression programs.

The background count matrix is derived from all barcodes captured in the experiment, specifically those with fewer than 100 UMIs, ensuring no cells are included. In a contaminated scRNA-seq experiment, the foreground barcodes primarily contain the native gene expression program A, the contaminated marker's gene expression program B, and the non-specific ambient gene expression program C (such as housekeeping genes, etc.). Programs B and C together determine the ambient expression in each sample. Program C exhibits a relatively uniform distribution and a minimal impact across most of the data, while the expression of Program B varies in each sample, primarily influenced by the number and diversity of contamination sources. If background barcodes exhibit a similar expression program to one of the foreground clusters, it implies that these background barcodes share a potential program B with their corresponding foreground clusters.

To identify background barcodes with expression programs similar to foreground clusters, two step Non-Negative Least Squares (NNLS) approximation is applied to background counts matrix:

$$H_{in} = \left[H_{i1}, H_{i2}, \ldots, H_{iq}\right], q = 1,2,\ldots n$$
$$Minimize: ||V_{kq} - W_{ki} \cdot H_{iq}||^2,\ H_{iq} \geq 0, q = 1,2,\ldots n$$

$$X_{in} = \left[X_{i1}, X_{i2}, \ldots, X_{iq}\right], q = 1,2,\ldots n$$
$$Minimize: ||H_{iq} - M_{ii} \cdot X_{iq}||^2,\ X_{iq} \geq 0, q = 1,2,\ldots n$$

In first step of NNLS, $V_{kn}$ is the unit variance scaled background counts matrix which has k features and n barcode. For each column of $V_{kn}$, which can be denoted as $V_{kq}$, NNLS seeks to find a non-negative coefficient vector $H_{iq}$ that best approximate $V_{kq}$ as a linear combination of the basis vectors in $W_{ki}$. $H_{iq}$ is calculated independently and finally forms the non-negative coefficient matrix $H_{in}$, which represents the distribution or the usage of each background barcode with respect to the ith program. In the second NNLS step, foreground cells within the same cluster are grouped together, and the median value of $H_{im}$ for each program is calculated within each cluster, denoted as $M_{ii}$. By minimizing the above equation for each column q independently, a non-negative coefficient matrix $X_{in}$ can be achieved, representing the contribution of each cluster expression program to each background barcode. Only background barcodes with a maximum proportion value larger than 0.95 are considered for subsequent steps, which means their expression programs closely match with one of the foreground clusters.

**Identification of potential seed ambient genes**

We next evaluate the ambient expression level and identify seed ambient genes using a smooth spline method with clusters sharing similar expression programs with at least one background barcode. As the ambient expression program has been divided into two types, we observed that cell type specific contamination program B show a discrete gap between background barcodes and their corresponding foreground clusters while non-specific contamination program C doesn't. The considerable difference between expected and actual gene expression compared to other genes will help to distinguish ambient genes in program B comparing to the others.

For each cluster c, we separately calculate average expression values for all genes g in background barcodes $d$ and foreground barcodes $r$ identified by the above step, denoted as $\mu^d_{g,c}$ and $\mu^r_{g,c}$ respectively. Each run of fitting involves randomly selecting genes and using the UnivariateSpline function in Python's scipy package to obtain the predicted average expression level of the genes in the background based on the foreground values. The average gene expressions in the foreground barcodes, are sorted in ascending order to ensure the data is an increasing vector.

$$fit_{g,c} = \text{UnivariateSpline}(\mu^d_{g,c}\ ,\ \mu^r_{g,c})\ (\mu^r_{g,c})$$

The distance of fitted values to actual average background expression can be denoted as:

$$D_{g,c} = fit_{g,c} - \mu^d_{g,c}$$

To ensure the stability of the model and the comprehensiveness of the fitting results, we use 60% of the genes for each fitting and perform 100 iterations. Genes with the highest $D_{g,c}$ are more likely to be potential ambient genes.

We then pooled the gene-$D_{g,c}$ pairs from each cluster together and the top 20 genes, sorted by $D_{g,c}$, are considered as potential seed ambient genes. Additionally, the average

value of the top 20 genes' $D_{g,c}$ is employed as the overall ambient level of the sample. A higher level signifies that the sample is severely affected by ambient expressions.

**Correction of ambient genes**

The final step is to capture all potential ambient genes based on the seed ambient genes according to their contribution to gene expression program $W_{ki}$ and find an ambient expression cutoff to correct them.

Potential ambient expression programs typically are from few source ambient clusters and contaminate other clusters. By examining the median coefficient value of barcodes from the $H_{im}$ matrix in each cluster, we can determine the weight of each program to the clusters. The program with the highest weight in the greatest number of clusters is considered the potential ambient expression program.

Next, genes with non-zero contribution to this ambient program are identified, and their Pearson's correlation with the previously identified seed ambient genes is used to capture more ambient genes. A gene is considered as a potential ambient gene if it has a correlation of 0.95 or higher with the previously identified seed ambient genes. The clusters where this ambient expression program has the highest weight are considered potential source ambient clusters.

Binary classification will then be used twice for each ambient gene: (1) identify all potential ambient source clusters (2) determine the ambient gene expression threshold.

Firstly, for each ambient gene, cells in the cluster with the highest average expression are assigned a label of 1. Cells in the remaining clusters are assigned a label of 0. We aim to use AUC-ROC here to determine if some clusters with a label of 0 cannot be distinguished from cells with label 1, thereby identifying additional potential source clusters of this particular ambient gene. If the obtained AUC-ROC value is less than or equal to 0.6, it indicates that the 0 and 1 clusters share a similar gene expression pattern.

$$\mathrm{AUC-ROC} = \int_0^1 TPR\ d(FPR)$$

TPR is short for the true positive rate, FPR is short for the false positive rate.

Secondly, we relabel the indistinguishable clusters identified above as 1, and binary classification can be performed again. Youden's J statistic is used to determine the best threshold of correctionfor each ambient gene.

$$\mathrm{Youden's\ J\ statistic} = \mathrm{TPR} - \mathrm{FPR}$$

By subtracting this count threshold from the expression values and ensuring non-negative values, the final corrected expression values are obtained.

**Existing Ambient RNA correction tools**

For CellBender (v0.3.0), remove-background module was used to run on the raw matrix, with the default recommended settings (fpr 0.01, epochs 150). The parameter expected-cells and total-droplets-included was set to 3,000 and 15,000.

For DecontX (v1.16.1) and SoupX (v1.6.2), raw matrix and filtered matrix were used as input. Cluster labels for filtered matrix were generated by Louvain (0.8.1) on 30 principal components and a resolution of 1 as implemented by FindClusters in Seurat (v4.3.0.1) after normalization and feature selection of 3000 genes. The ambient RNA correction processes for DecontX and SoupX both utilized default parameters.

**Processing of mouse kidney and human PBMC scRNA-seq data**

The public kidney scRNA-seq data, with annotation and genotype estimate information are available at https://zenodo.org/records/7328632. For these three kidney samples, we removed the top 2% of barcodes with the highest UMI counts and the top 2% with the highest unique gene counts for outliers. We also excluded barcodes with a mitochondrial read fraction beyond 50%, and gene counts lower than 200. Dimensionality reduction and clustering analysis were performed using Scanpy (v1.9.3) [12].

The public PBMC scRNA-seq data from 10x Genomics, analyzed with CellRanger version 2.1.0, is available at https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0. For this sample, following the methods described previously[5], we removed the top 5% of barcodes with the highest UMI counts and the top 5% with the highest unique gene counts. We also excluded barcodes with a mitochondrial read fraction beyond 10%. Dimensionality reduction and clustering analysis were performed using Scanpy (v1.9.3) [12].

**Supplementary Information**

**Additional file 1:**

Figure. S1. Performance evaluation of existing tools on mouse kidney samples.

Figure. S2. Performance evaluation of CellClear and CellBender on a human PBMC sample.

**Additional file 2:**

**Table S1.** Each run by CellClear will output ambient expression level and ambient genes.

**Declarations**

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

CellClear and the code used in this study is available at https://github.com/WhiteRabBio/CellClear under the BSD 3-Clause License.

Competing interests

All authors are current or former employees of Singleron Biotechnologies.

Funding

This work was supported by the National Science Foundation of Jiangsu Province (BK20230278).

Acknowledgements

Not applicable

## References


[1]. Wang, Yong, and Nicholas E Navin. “Advances and applications of single-cell sequencing technologies.” *Molecular cell* vol. 58,4 (2015): 598-609. doi:10.1016/j.molcel.2015.05.005

[2]. Liu, Yi-Dan et al. “Deciphering Adult Neural Stem Cells with Single-Cell Sequencing.” *Stem cells and development* vol. 32,9-10 (2023): 213-224. doi:10.1089/scd.2022.0234

[3]. Brennecke, P., Anders, S., Kim, J. et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10, 1093–1095 (2013). https://doi.org/10.1038/nmeth.2645

[4]. Yang, Shiyi et al. “Decontamination of ambient RNA in single-cell RNA-seq with DecontX.” *Genome biology* vol. 21,1 57. 5 Mar. 2020, doi:10.1186/s13059-020-1950-6

[5]. Fleming, Stephen J et al. “Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender.” *Nature methods* vol. 20,9 (2023): 1323-1335. doi:10.1038/s41592-023-01943-7

[6]. Young, Matthew D, and Sam Behjati. “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data.” *GigaScience* vol. 9,12 (2020): giaa151. doi:10.1093/gigascience/giaa151

[7]. Berg, Marijn et al. “FastCAR: fast correction for ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets.” *BMC genomics* vol. 24,1 722. 29 Nov. 2023, doi:10.1186/s12864-023-09822-3

[8]. Janssen, Philipp et al. “The effect of background noise and its removal on the analysis of single-cell expression data.” *Genome biology* vol. 24,1 140. 19 Jun. 2023, doi:10.1186/s13059-023-02978-x

[9]. Yu, Guangchuang et al. “clusterProfiler: an R package for comparing biological themes among gene clusters.” *Omics : a journal of integrative biology* vol. 16,5 (2012): 284-7. doi:10.1089/omi.2011.0118

[10]. Efremova, Mirjana et al. “CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.” *Nature protocols* vol. 15,4 (2020): 1484-1506. doi:10.1038/s41596-020-0292-x

[11]. Nijmeijer, Saskia et al. “Adhesion GPCRs in immunology.” *Biochemical pharmacology* vol. 114 (2016): 88-102. doi:10.1016/j.bcp.2016.04.013

[12]. Wolf, F Alexander et al. “SCANPY: large-scale single-cell gene expression data analysis.” *Genome biology* vol. 19,1 15. 6 Feb. 2018, doi:10.1186/s13059-017-1382-0

[13]. Elosua-Bayes, Marc et al. “SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.” *Nucleic acids research* vol. 49,9 (2021): e50. doi:10.1093/nar/gkab043