

Methodological exploration of cross-platform single-cell comparison and joint exploratory analysis of IgG4-ROD

Zhihao Zhao¹, Pei Liu¹, Yu Deng¹, Haibei Dong², Zhiyang Zhang³

1. School of Life Science and Food Engineering, Jiangsu Normal University, Xuzhou; 2. Department of Oncology, Xuzhou First People's Hospital, Xuzhou; 3. Department of Ophthalmology, Xuzhou First People's Hospital, Xuzhou

Abstract | Using the same reproducible pipeline, this study examined how merging and filtering strategies in cross-platform single-cell comparison alter functional enrichment conclusions, and accordingly proposed actionable practical recommendations. Cases were 10x data from four IgG4-ROD lacrimal glands, and controls were three SORT-seq technical plates of healthy lacrimal glands from the same donor. Because platform and disease were completely confounded and the controls were technical replicates, this article reports only methodological conclusions and makes no biological inferences. Methodologically, we defined four differential gene criteria from broad to strict. We used within-sample proportions as the analysis unit, exhaustively enumerated all 35 possible exact permutations, and applied Benjamini–Hochberg (BH) correction. Because only processed matrices were available for controls, we used sample-level permutation instead of DESeq2. We also performed doublet detection, ambient RNA quantification, and linear correction. Main findings: before merging and without taking a gene intersection, 53.7% (9,254/17,241) of genes were detected in fewer than 5% of cells in at least one group, and enrichment was dominated by cytoskeletal/neural/ion-related terms. After taking the intersection and filtering, enrichment shifted to RNA splicing and immune-related terms, suggesting that the filtering criterion was more critical than filtering stringency. Proportion tests showed no significant between-group differences in permutation tests, with the smallest BH-adjusted value of 0.1286. For single-gene sample-level permutation, the smallest attainable P was 0.0556, and the smallest BH q was 0.133 (above the lower limit by 0.0556). Beyond the single-cell data, serum miRNA analysis was also incorporated into the joint exploratory analysis (methodological replication only, not a biological finding); under the primary criterion (excluding zero-variance probes) all 15 miRNAs were non-significant (BH 0.0651–0.0711); the all-probe value BH P = 0.0476 reflects step-up accumulation of same-order raw P values and is not presented as a finding.

Keywords | IgG4-related ophthalmic disease; single-cell RNA sequencing; cross-platform integration; gene filtering criteria; ambient RNA; serum miRNA

Copyright © 2026 by the author(s).

This article is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

<https://creativecommons.org/licenses/by-nc/4.0/>



Correspondence: Zhiyang Zhang (yuerlili@163.com); Haibei Dong (donghaibei0516@163.com)

These authors contributed equally: Zhihao Zhao, Pei Liu

<https://doi.org/10.68070/jyq03911>

Introduction

IgG4-related disease (IgG4-RD) is a chronic immune-mediated disease that can involve the lacrimal glands, salivary glands, and pancreas. Approximately one third of patients develop ocular involvement (IgG4-ROD). The lacrimal gland is the most commonly involved target organ. Its diagnosis and treatment follow the Chinese expert consensus [1]. Single-cell and organoid studies of the human lacrimal gland have been reported [2]. Clinically, IgG4-ROD must be differentiated from lymphoma and Sjögren's syndrome, and diagnosis depends on biopsy. Glucocorticoids often induce initial remission, but efficacy decreases after fibrosis develops [3]. Therefore, identifying immune microenvironment changes before fibrosis has clinical value. Previous mechanistic studies have mainly focused on CD4⁺ T cells, especially Tfh cells, in IgG4 class switching [4]. However, the magnitude of local cell-composition changes in the lacrimal gland and the true ranking of pathways have not been systematically characterized within a single analytical pipeline.

The sample conditions in this study raised two technical issues. First, case samples were generated with 10x and control samples with SORT-seq, so platform and disease status were completely confounded. In addition, controls came from only three technical plates from one donor. Second, if no gene intersection was taken before merging, the merge step would assign zero values on one platform to genes detected only on the other platform. This would produce many false differentially expressed genes with zero detection in one group. Both library-preparation types also contained ambient RNA contamination, which could cause false downregulation and false upregulation, respectively.

Accordingly, this study used the same reproducible pipeline to answer three methodological questions: Q1, whether taking a gene intersection and applying detection-rate filtering before merging alters functional enrichment conclusions; Q2, whether threshold-based enrichment and whole-gene ranking GSEA yield consistent conclusions; and Q3, how the differential gene profile and within-cell-type signals change before and after ambient RNA correction. In addition, using MHC class II genes as an example, we demonstrate how the within-cell-type workflow is implemented and explain why its results cannot be

extrapolated across cell types. This study does not pursue new biological conclusions but aims at reusable analytical specifications; because platform and disease were completely confounded and the controls were technical replicates, all between-group comparisons are used only for methodological inference.

1 Materials and Methods

1.1 Data acquisition and processing.

Raw 10x data from the four cases were downloaded from the European Nucleotide Archive (ENA; accession numbers SRR33333120–SRR33333123). GEO and other public platforms did not provide directly usable processed expression matrices for this project; therefore, we performed alignment and counting ourselves. The control dataset GSE164403 and the serum miRNA dataset GSE155177 were downloaded from Gene Expression Omnibus. For GSE164403, the provided processed expression matrix and metadata were used (the Sort field annotates the sorting method; in this study, 1,152 cells with Sort=All from donor 3 were used) [2]. For GSE155177, the miRNA expression profiles and group information were used. All data cleaning, integration, quality control, and statistics were performed in R (R 4.5.1). Quality control used Seurat's PercentageFeatureSet and platform-specific [5,6] thresholds for nFeature_RNA/nCount_RNA/percent.mt: 10x data used 200–8,000 genes, >500 UMI, and mitochondrial percentage <15%; SORT-seq data used 500–10,000 genes, >1,500 UMI, and mitochondrial percentage <20%. Batch effects between platforms were corrected with Harmony on 30 principal components, followed by neighbor graph construction and UMAP. Differential expression, enrichment, and statistics used Seurat, clusterProfiler, fgsea [7], and limma, respectively.

1.2 Data sources and ethics

The case group consisted of 10x data from four pathologically confirmed IgG4-ROD lacrimal glands (ENA: SRR33333120–SRR33333123) [8]. Controls were three plates from GSE164403 healthy lacrimal glands from the same donor (Sort=All; no specific cell sorting, all viable cells retained). Serum miRNA data were from GSE155177 (11 patients with IgG4-ROD and 11 healthy controls, GPL21263). Public de-identified data were used; therefore, no additional ethics approval was required. Ethics approval details for the original studies are reported in the original publications [2,3,9]. Both case and control data were

public de-identified data. The original articles did not provide detailed clinical phenotypes such as age, sex, disease duration, or serum IgG4 levels; therefore, clinical association analysis could not be performed. Heterogeneity among the four cases was substantial and is described in Section 2.2.

1.3 Data preprocessing, integration, and cell annotation

FASTQ files were processed with fastp to remove adapters and low-quality reads (-A, --disable_length_filtering, -l 0, -w 4). After filtering, 3.15×10^9 reads remained, with Q30 of 73.4%–75.3%. Reads were aligned and counted with Cell Ranger count against a custom Ensembl release 113 GRCh38 primary assembly. The FASTA and GTF files were obtained from https://ftp.ensembl.org/pub/release-113/fasta/homo_sapiens/, and the annotation was GENCODE-compatible. The reference contained 44,757 genes, including lncRNAs (--nosecondary --create-bam=false). The estimated total cell number was 42,945. This reference contained more genes than the standard 10x reference of 36,601 genes.

1.4 Differential expression, filtering criteria, and functional enrichment

Before merging, 17,241 common genes were retained. 10x data were log-normalized, and SORT-seq data used the original log1p values. A total of 2,000 highly variable genes were selected, scaled after regressing out mitochondrial percentage by dataset, and integrated with Harmony [10] using 30 principal components. A neighbor graph and UMAP were then constructed at resolution 0.5. After platform-specific quality control, 42,457 cells were retained: 41,680 cases and 777 controls.

1.5 Within-cell-type analysis, robustness tests, and miRNA analysis

Following single-cell analysis best practices [11], cells were annotated into 11 types based on marker genes: acinar (LCN1/LTF/LYZ), ductal (KRT7/KRT19), myoepithelial (KRT5), fibroblast (DCN), endothelial (PECAM1/VWF), and immune cells (IL7R/CD8A/NKG7/MS4A1/MZB1, etc.). Some markers had limited specificity. For example, LYZ was also expressed in acinar cells, and IL7R and NKG7 were also detected in other lymphocyte populations. Therefore, cell-type annotation may have involved type confusion; labels were used only for stratification and not as independent conclusions.

Given the risk of pseudo-replication in single-cell differential expression [12], within-sample proportions were used as the analysis unit. Because there were only four cases and three technical plate controls, exact permutation testing was used. All $C(7,4)=35$ possible groupings of the seven samples were exhaustively enumerated. Between-group rank-sum differences were calculated. Two-sided P was defined as $2 \times \min(\text{one-sided})$. The minimum attainable P was 0.0571. Benjamini–Hochberg correction [13] results and bootstrap 95% confidence intervals (5,000 times, resample based on the sample) were reported; the latter were used only to describe the range of effect sizes. The minimum attainable values differed between the two two-sided definitions: 0.0571 and 0.0556. In this study, the former was used for cell-level proportion tests, and the latter for sample-level and single-gene permutation tests; both are greater than 0.05, so neither family of tests can reach significance at any effect size, and their results are interpreted only as consistency checks. Sample-level permutation involved 14 genes $\times$ 4 cell types. Because controls provided only processed matrices and no raw counts, edgeR/DESeq2 could not be used, so sample-level permutation was used as an alternative. After additional BH correction, the smallest q value among 56 single-gene $\times$ cell-type tests was 0.133, and no test had $q < 0.05$. Results were uniformly reported as “reached the smallest value attainable under this design ($P = 0.0556$) and therefore cannot be interpreted as evidence of a difference.”

Differential expression used FindMarkers with the Wilcoxon test and Benjamini–Hochberg correction [13], with logfc.threshold=0, covering 17,241 genes. Four criteria were defined: A was adjusted $P < 0.05$ (9,473 genes); B was A plus detection rates $\geq 10\%$ in both groups (3,095 genes); C was B after removing technical genes (2,981 genes); D (primary criterion) was C plus $|\log_2\text{FC}| > 0.5$ (2,538 genes). Cell-level P values had pseudo-replication and were used only for candidate ranking. Differential expression was based on the data layer of the RNA assay, which contains log-normalized data; SCT or Harmony embedding space was not used. Cross-platform comparison was performed only on common genes using this layer. Technical genes refer to genes whose expression is mainly determined by sequencing depth, platform capture efficiency, or high-abundance transcripts rather than by the biological state of interest; mitochondrial gene percentage is a routine quality-control metric [11],

and high-abundance transcripts are also a major source of ambient RNA contamination [14]. Accordingly, this set was constructed in-house using gene-symbol prefix rules (not imported from an external database; the rules were MT-, RPS/RPL, MRPS/MRPL, and EEF1 prefixes, plus six specific symbols: MALAT1, B2M, TMSB4X, FAU, UBA52, and RACK1), comprising 212 genes (1.23% of the 17,241 tested genes) and covering four categories: mitochondrially encoded genes, mitochondrial and cytoplasmic ribosomal proteins, and high-abundance housekeeping/translation-related transcripts; hemoglobin genes were not included. Of these, 114 genes that also passed criteria A and B were removed under criterion C. The rationale for defining criterion D as the primary criterion was as follows. After cross-platform merging, 53.7% (9,254/17,241) of the tested genes had a detection rate <5% in at least one group. Without detection-rate and technical-gene filtering, enrichment results would be dominated by cytoskeletal/neural/ion-related terms. Therefore, criterion D, which simultaneously required statistical significance, an effect-size threshold, and detection in both groups, was used as the interpretable gene pool. Complete gene lists for each criterion, one-sided detection-rate distributions, and complete enriched pathway lists are provided in the intermediate analysis results and are available from the corresponding author upon request.

Threshold-based enrichment was performed on gene lists from each criterion using GO-BP enrichment with clusterProfiler [15], BH correction, and $P < 0.05$. Whole-gene ranking enrichment used gseGO with fgsea [7,16], minGSSize=15, and maxGSSize=500.

Pseudobulk mean expression was calculated for each sample $\times$ cell type combination with at least 10 cells. Because the control dataset provided only processed expression matrices and no raw counts, edgeR/DESeq2 could not be used. Therefore, sample-level permutation testing was used to evaluate overall shifts and avoid pseudo-replication. A leave-one-donor-out sensitivity analysis was also performed. The statistic was defined as follows: for each gene g , $\Delta g = \text{mean}(\text{case expression}) - \text{mean}(\text{control expression})$, and then $T = \text{mean}(\Delta g)$. The null distribution was obtained by exhaustively recalculating T under all $C(7,4)=35$ case/control groupings. Two-sided P used add-one correction, $P=(k+1)/36$, and the minimum attainable P was $2/36=0.0556$. Cell-level proportion tests used

$2 \times \min(\text{one-sided})$ without add-one correction, with a minimum attainable P of $2/35=0.0571$. In sensitivity analysis, the median $\log_2\text{FC}$ of each gene was used instead of the mean as the statistic; the direction of conclusions did not change, and all results were non-significant. Permutation was performed independently within each cell type, and the 53 sample $\times$ cell type combinations were not combined for permutation.

Doublet detection was performed only on the four 10x case samples. Artificial doublets were generated by random pairing. After merging, PCA was performed, and doublet scores were calculated with kNN at $K=20$. The threshold was determined according to the expected rate from the official 10x curve.

To assess whether Harmony erased disease signals, independent clustering was performed by platform. Majority Harmony label transfer within each cluster was then used, and type consistency and direction consistency of proportion differences were compared. For each cell type $\times$ ambient marker gene set combination, mean expression and detection rate were calculated to quantify background contamination.

Limma [17] was used for analysis with the model $\sim 0 + \text{group}$ and the contrast Case-Control, and probes were mapped to standard miRBase names. The threshold was $|\log_2\text{FC}| \geq 1$ and adjusted $P < 0.05$. Two criteria were reported: all probes (2,566) and excluding zero-variance probes (2,029), with the latter as the primary criterion.

Predicted target genes of differential miRNAs were obtained from miRDB [18]. RefSeq transcript numbers were mapped to gene symbols and deduplicated, yielding 4,621 genes. These were intersected with the 2,538 interpretable differential genes from criterion D. A hypergeometric test with the upper tail was performed with 17,241 tested genes as background. A 10,000-iteration random sampling permutation was also performed to account for gene dependence.

2 Results and Analysis

Sections 2.1–2.3 present the main methodological results. Signals within cell types are exploratory descriptions and are not core findings.

2.1 Quality control and cell atlas

For the four case samples, 4.25×10^9 reads were obtained before filtering. After fastp, 3.15×10^9 reads

were retained, and Q30 increased from 68.8%–73.9% to 73.4%–75.3%. Cell Ranger estimated a total of 42,945 cells and after merging with controls and selecting common genes, there were a total of 43,721 cells. After platform-specific quality control, 42,457 cells were retained: 41,680 cases and 777 controls (Figure 1A, Table 1).

After Harmony integration and clustering at resolution 0.5, 23 clusters were annotated into 11 cell types: acinar cells, ductal epithelial cells, myoepithelial cells, fibroblasts, endothelial cells, CD4⁺ T cells, CD8⁺ T cells, NK cells, B cells, plasma cells, and dendritic cells (Figure 1B, 1C). After integration, cells from the two platforms were no longer completely separated in the shared space, but residual platform effects remained visible. Therefore, integration was

used only to establish a comparable space; we do not guarantee that platform effects were completely removed.

Regarding the s008 plate, we analyzed this plate from donors 7 and 8. All 384 cells showed an acinar secretory phenotype, so this plate could not serve as a whole-cell control and was not included in the main analysis. In a separate acinar cell-restricted comparison, acinar secretory genes had consistent detection rates in the two healthy donors but were markedly lower in cases (PRH1 99%/100% vs 21%). MHC class II genes were lower in the healthy samples and higher in the case samples, consistent with the direction observed in the main analysis (15/15). This result is recorded only for direction and is not used for between-group inference

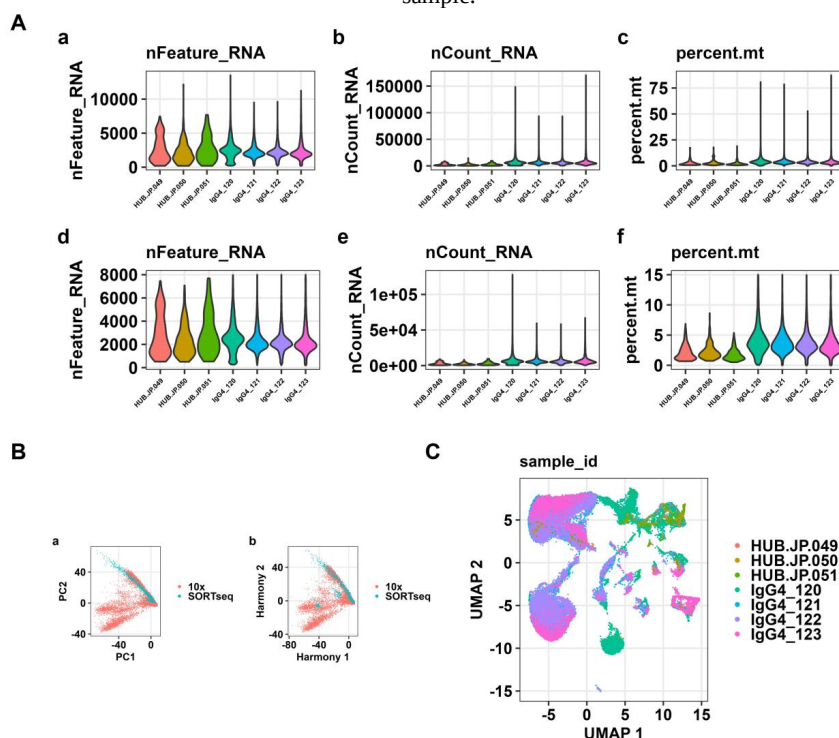
Table 1 Sequencing and quality-control metrics for each sample: raw barcode number, CellRanger-estimated cell number, cell number after each filtering step, median UMI, median gene number, median mitochondrial percentage, and doublet detection rate (observed vs expected). All values in the table are descriptive results, and no between-group inference was performed (see Methods 1.4).

Sample	Raw barcodes / wells	Cells estimated by Cell Ranger	Cells before QC (after merging)	Cells after QC	QC retention rate	Median UMI per cell	Median genes per cell	Median mitochondrial fraction	Doublets detected (observed vs expected)
IgG4_120 (SRR33333120)	2,343,992	11,867	11,866	11,168	94.1%	6,051	2,466	3.72%	1,015(9.09% vs 8.42%)
IgG4_121 (SRR33333121)	1,917,434	14,816	14,801	14,578	98.5%	5,697	2,131	3.48%	1,720(11.80% vs 10.80%)
IgG4_122 (SRR33333122)	2,883,509	10,404	10,393	10,294	99.1%	5,348	2,118	3.37%	917(8.91% vs 7.81%)
IgG4_123 (SRR33333123)	1,352,996	5,858	5,799	5,640	97.3%	5,715	2,064	3.16%	250(4.43% vs 4.41%)
Case total	8,497,931	42,945	42,859	41,680	97.3%	—	—	—	3,902(9.36% vs 8.56%)
HUB.JP.049	384	—	312	283	90.7%	2,147*	2,374	1.80%	Not performed (SORT-seq)
HUB.JP.050	384	—	282	250	88.7%	1,812*	2,000	1.97%	Not performed (SORT-seq)
HUB.JP.051	384	—	268	244	91.0%	2,368*	2,521	1.56%	Not performed (SORT-seq)

Table 1.(continue)Sequencing and quality-control metrics for each sample: raw barcode number, CellRanger-estimated cell number, cell number after each filtering step, median UMI, median gene number, median mitochondrial percentage, and doublet detection rate (observed vs expected). All values in the table are descriptive results, and no between-group inference was performed (see Methods 1.4).

Sample	Raw barcodes / wells	Cells estimated by CellRange r	Cells before QC (after merging)	Cells after QC	QC retention rate	Median UMI per cell	Median genes per cell	Median mitoch ondrial fraction	Doublets detected (observed vs expected)
Control total	1,152	—	862	777	90.1%	—	—	—	—
Total	—	42,945	43,721	42,457	—	—	—	—	3,902

Figure 1. Quality control and integration. (A) Quality-control metrics for each sample (each point represents one sample; four cases and three control plates); (B) PCA before (left) and after (right) integration, colored by platform; (C) UMAP colored by sample.



2.2 Trend observations of cell composition and immune infiltration

Proportions of each cell type are shown in Figure 2A and 2B, and specific values are shown in Table 2. Using within-sample proportions as the analysis unit, the proportions of CD4+ T cells, B cells, plasma cells, and endothelial cells showed increasing trends in the case group. The proportions of acinar cells,

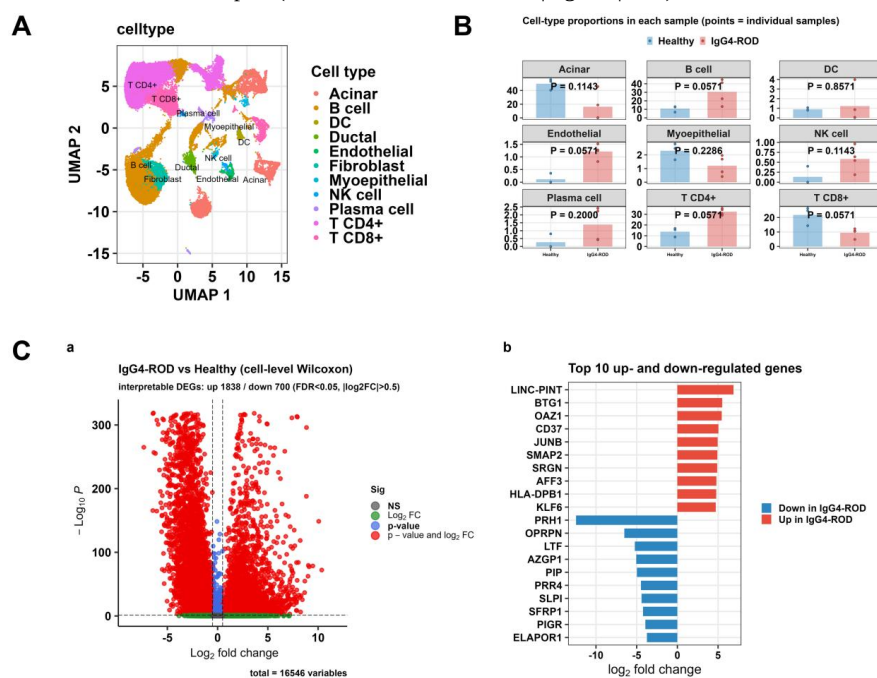
myoepithelial cells, and CD8+ T cells showed decreasing trends. After exact permutation testing and BH correction [13], all comparisons reached the lower limit of this design but remained >0.05 (minimum adjusted $P=0.1286$). Because controls came from only three technical plates from one donor, and ductal epithelial cells and fibroblasts were 0 cells in controls.

Table 2. Proportions of each cell type in lacrimal glands in the two groups (mean within-sample proportion, %) and exact permutation test with Benjamini–Hochberg correction^[13].

All values in the table are descriptive results, and no between-group inference was performed (see Methods 1.4).

Cell type	IgG4-ROD (%)	Healthy control (%)
Acinar cells	16.29	49.84
B cells	30.43	10.88
CD4 ⁺ T cells	32.36	13.75
CD8 ⁺ T cells	9.55	21.83
Endothelial cells	1.21	0.12
Myoepithelial cells	1.20	2.30
NK cells	0.58	0.13
Plasma cells	1.39	0.27
Dendritic cells (DC)	1.23	0.89

Note: Because single-cell differential expression has a pseudo-replication risk^[12], within-sample proportions were used as the analysis unit: four cases and three technical plates from one donor as controls. Two-sided P was $2 \times \min(\text{one-sided})$, with a minimum attainable value of 0.0571. Ductal epithelial cells and fibroblasts were 0 cells in controls and were not comparable, so they are not listed. These control data cannot represent lacrimal gland ductal or stromal components; therefore, between-group comparisons involving these two categories are invalid. Proportions in the table are group means of within-sample proportions. Because only major types are listed, they do not sum to 100%.

Figure 2. Cell types and differential expression. (A) UMAP of 11 cell types (23 clusters, resolution 0.5); (B) within-sample proportions (each point represents one sample; exact permutation C(7,4)=35, minimum BH-adjusted value 0.1286); (C) volcano plot (Wilcoxon, BH correction, $|\log_2 \text{FC}| > 0.5$).

2.3 Methodological validation of differential expression, enrichment, and miRNA (miRNA: methodological replication only)

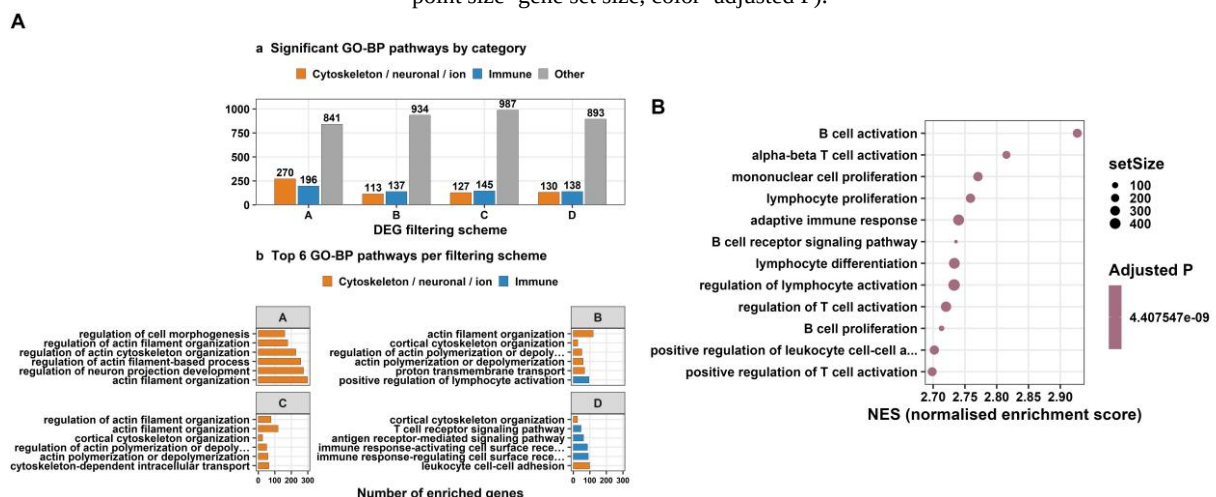
When merging without taking a gene intersection or applying detection-rate filtering, 9,254 (53.7%) of the 17,241 tested genes had a detection rate <5% in at least one group. After applying significance filtering, criterion A yielded 9,473 significant genes, of which 4,841 (51.1%) also belonged to the aforementioned 9,254 genes with insufficient detection; as the criteria were tightened, the number of genes entering enrichment decreased successively to 3,095 (B), 2,981 (C), and 2,538 (D). Stepwise decomposition showed that the contributions of the three filtering steps to gene-number reduction were 92.0% (detection rate), 1.6% (technical genes), and 6.4% ($|\log_2FC| > 0.5$), respectively, indicating that detection-rate filtering dominated the contraction of the gene pool. In this study, significant GO-BP pathways were classified by keyword in term names into three categories: cytoskeletal/neural/ion-related (neural, axon, synapse, etc.), immune-related (immune, leukocyte, lymphocyte, etc.), and other. Among the 1,307 significant pathways under criterion A, 270 (20.7%) were cytoskeletal/neural/ion-related and 196 (15.0%) were immune-related; among the 1,161 significant pathways under criterion D, the two categories comprised 130 (11.2%) and 138 (11.9%) pathways, respectively, and cytoskeletal/neural/ion-related terms changed from outnumbering immune-related terms to being slightly fewer than them (Figure 3A). Fisher's exact test showed that the decrease in the proportion of

cytoskeletal/neural/ion-related terms between A and D was significant ($OR=2.07$, $P=1.7 \times 10^{-10}$), whereas the decrease for immune-related terms was smaller ($OR=1.31$, $P=0.025$). This suggests that merging without taking an intersection systematically produces false-positive pathways dominated by cytoskeletal/neural/ion-related terms.

Composition of differential genes: Among the 2,538 interpretable differential genes under criterion D, 1,838 were upregulated and 700 were downregulated (Figure 2C). Upregulated genes were mainly MHC class II genes, such as HLA-DPB1 and HLA-DRA, and genes related to B/T cell receptor signaling. Downregulated genes were mainly acinar secretory function-related genes, such as PRH1 and LTF. The complete gene list is provided in the supplementary materials.

To exclude the possibility that upregulation was only caused by increased immune cell proportions, we further compared within cell types. MHC class II signals, including HLA-DPB1 and HLA-DRA, showed positive descriptive shifts on the case side in acinar and B cells. In T cells, these signals were greatly attenuated after simplified ambient RNA correction, suggesting possible ambient RNA effects. In addition, the top enriched terms within each type were cytoplasmic translation and ribosome-related pathways, suggesting that platform capture differences dominated pathway-level results within cell types (Figure 3B). Values in this section and within-cell-type FindMarkers are based on different normalization bases, so only directions can be compared, not absolute values.

Figure 3. Effect of filtering criteria on enrichment. (A) Counts of significant GO-BP pathways by category under the four criteria (orange=cytoskeletal/neural/ion, blue=immune, gray=other) and the top 6 terms for each criterion; (B) GSEA (fgsea, x-axis NES, point size=gene set size, color=adjusted P).



Whole-gene ranking results are shown in Figure 3B. The most significant threshold-based enrichment term was RNA splicing, with an extremely small adjusted P value. The ranking from whole-gene GSEA was clearly different; almost all top NES terms were immune activation terms, including B cell activation at 2.94 and $\alpha\beta$ T cell activation at 2.83 (Figure 4A, 4B). Quantitative comparison showed that there was no overlap between the two methods for Top10 or Top20 pathways (only 1 pathway for Top50 and 12 for Top100), and the Spearman rank correlation of rankings across 335 identically named pathways was only 0.13, indicating that the pathway rankings of the two methods were essentially independent. The difference between the two methods is itself one of the conclusions. Combined with within-cell-type results, we limit functional conclusions to directional indications.

Robustness checks: sample-level permutation and pseudobulk. The pseudobulk matrix was 17,241 genes $\times$ 53 combinations. Sample-level exact permutation showed that the overall shift directions of CD4⁺ T cells, CD8⁺ T cells, B cells, and acinar cells were consistent with the proportion tests, but none were significant, with a minimum attainable $P=0.0556$. Cell-level and pseudobulk log2FC were highly consistent, with Spearman 0.861 and direction consistency 99.9%. After leave-one-donor-out, pathway ranking Spearman correlation was ≥ 0.992 . The top pseudobulk pathway was also cytoplasmic translation. Regarding the combination number, “53 combinations” refers to the total number of sample $\times$ cell type combinations formed by seven samples and cell types with ≥ 10 cells in that sample. It is not simply 4 cell types $\times$ 7 samples = 28, because 11 cell types were included in the analysis, and 53 combinations remained after applying the ≥ 10 cells per combination threshold.

Technical checks: doublets and ambient RNA. Doublet detection used an in-house pipeline (artificial doublets $\rightarrow$ PCA $\rightarrow$ kNN scoring, with the same

principle as Scrublet/DoubletFinder [19,20]) and was applied only to the four 10x samples. The detection rate in each sample was 4.43%–11.80%, consistent with the expected multiplet rate, totaling 3,902 doublets (9.36%; Table 1). After removal, the maximum change in the proportion of each cell type was only 0.43 percentage points, and the log2FC changes of 12 key genes were all ≤ 0.03 with consistent directions; therefore, doublets were not removed from the main analysis and were reported only as a sensitivity analysis. Ambient RNA correction followed the SoupX approach [14,21]. The estimated contamination proportions for each cell type were 0.39–0.77: after correction, the positive shifts of HLA-DPB1/DRA in acinar and B cells were further enhanced (1.60 $\rightarrow$ 1.67, 1.84 $\rightarrow$ 2.32; 2.31 $\rightarrow$ 2.81, 2.72 $\rightarrow$ 3.48), whereas they were markedly attenuated in T cells (CD4⁺ T: 0.26 $\rightarrow$ 0.06, 0.25 $\rightarrow$ 0.02; CD8⁺ T: 0.76 $\rightarrow$ 0.38), suggesting that the MHC II signal in the latter may be mainly affected by ambient RNA and that between-group differences cannot be inferred from it. The median type consistency between independent clustering and the Harmony scheme was 0.977, and the group direction consistency rate was 90.9%.

Methodological replication of serum miRNA. Serum miRNA analysis was performed for methodological replication only and does not constitute a biological finding: under the primary criterion excluding zero-variance probes, none of the 15 miRNAs was significant (BH 0.0651–0.0711); under the all-probe criterion, BH-adjusted [13] P was 0.0476 (Table 3, Figure 4C, 4D), an artifact of step-up correction on same-order raw P values and therefore not treated as significant evidence. Its predicted target genes (4,621) overlapped with the interpretable differential genes (2,538) by 772 genes, with an expected overlap of 680.2 genes; the enrichment fold was only 1.135 (hypergeometric $P=6\times 10^{-6}$, permutation $P=1\times 10^{-4}$), which also does not constitute a biological conclusion.

Table 3. Exemplary differentially expressed miRNAs (methodological replication only, not a biological finding; all-probe criterion) and Benjamini–Hochberg corrected [13] P values under two probe criteria.

miRNA	log2FC	Raw P	BH-adjusted P (all probes, n=2,566)	BH-adjusted P (zero-variance removed, n=2,029)
hsa-miR-550a-5p	1.544	4.45×10^{-5}	0.0476	0.0651
hsa-miR-642b-3p	1.449	4.86×10^{-5}	0.0476	0.0651

Table 3.(continue) Exemplary differentially expressed miRNAs (methodological replication only, not a biological finding; all-probe criterion) and Benjamini–Hochberg corrected ^[13] P values under two probe criteria.

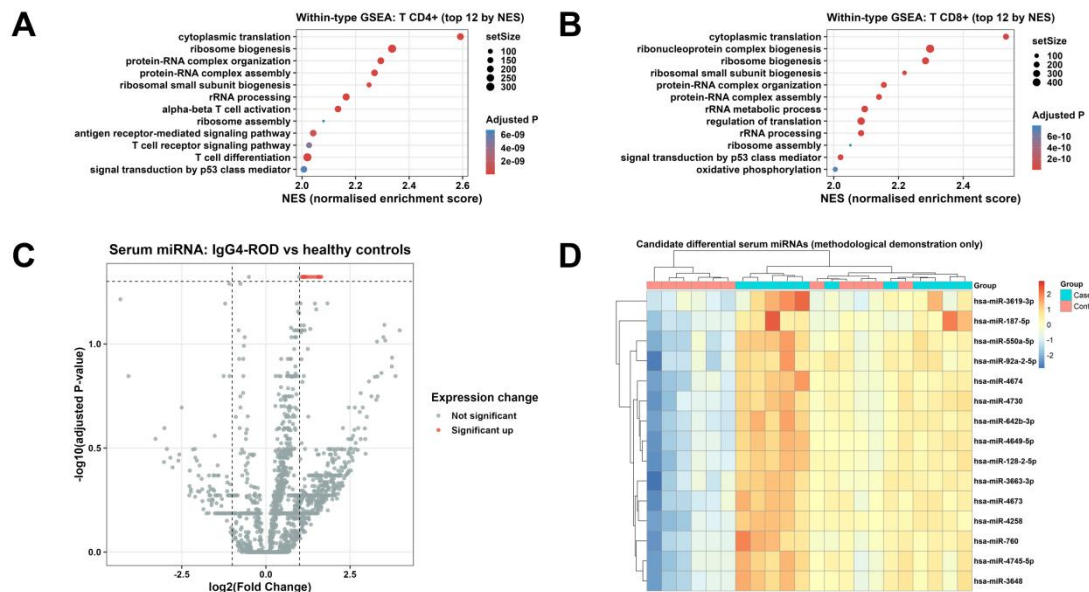
miRNA	log2FC	Raw P	BH-adjusted P (all probes, n=2,566)	BH-adjusted P (zero-variance removed, n=2,029)
hsa-miR-4730	1.571	7.97×10^{-5}	0.0476	0.0651
hsa-miR-4673	1.356	8.68×10^{-5}	0.0476	0.0651
hsa-miR-4649-5 p	1.193	1.62×10^{-4}	0.0476	0.0651
hsa-miR-128-2-5 p	1.265	1.82×10^{-4}	0.0476	0.0651
hsa-miR-4745-5 p	1.592	1.84×10^{-4}	0.0476	0.0651
hsa-miR-92a-2-5 p	1.540	2.10×10^{-4}	0.0476	0.0651
hsa-miR-3663-3 p	1.152	2.15×10^{-4}	0.0476	0.0711
hsa-miR-3648	1.646	2.29×10^{-4}	0.0476	0.0651
hsa-miR-4674	1.156	2.60×10^{-4}	0.0476	0.0711
hsa-miR-4258	1.070	2.73×10^{-4}	0.0476	0.0711
hsa-miR-187-5p	1.093	2.84×10^{-4}	0.0476	0.0711
hsa-miR-3619-3 p	1.122	2.98×10^{-4}	0.0476	0.0711
hsa-miR-760	1.134	3.15×10^{-4}	0.0476	0.0711

Note: Under the all-probe criterion with 2,566 probes, the raw P values of the 15 miRNAs were of the same order of magnitude. The step-up cumulative minimum made the adjusted P values all 0.0476. Under the primary criterion excluding zero-variance probes with 2,029 probes, values were 0.0651–0.0711, all non-significant. The two BH-adjusted P columns came from two independently run limma models with different prior variances and ranks, so they are not proportional reductions.

Complete statistics for the target gene intersection, including the number of overlapping genes, expected value, and hypergeometric/permutation tests, are provided in the supplementary materials. Because genes are strongly dependent, the hypergeometric test

may overestimate significance, so permutation results were used as the reference. The intersection enrichment fold was very small and cannot be used as a diagnostic biomarker or core finding. It is only a methodological example.

Figure 4. (A, B) GSEA within two T cell subsets, CD4⁺ T and CD8⁺ T, method demonstration only; (C) miRNA volcano plot; (D) miRNA heatmap methodological replication only, not a biological finding; primary criterion BH 0.0651–0.0711, all non-significant; all-probe BH 0.0476.



3 Discussion

This study has several fundamental limitations. First, controls came from only three technical plates from one donor and were technical replicates rather than independent biological replicates. The entire dataset contained only four donors, and only this one donor was unsorted. The other two donors, PatientNumber 7 and 8, were present only in EPCAM⁺ enrichment plates with 381 epithelial cells. They were not included because the sorting method differed and the cell number was insufficient. Second, platform and disease were completely confounded: all cases were 10x and all controls were SORT-seq, so effects could not be separated. Third, there were only 777 control cells, representing 1.8% of cells, and heterogeneity among the four cases was substantial, so statistical power was very low. With $C(7,4)=35$ permutations, the minimum attainable P was $0.0571 > 0.05$, so no effect size can reach significance at $\alpha=0.05$; the observed largest comparable difference (acinar cells 16.29% vs 49.84%, a difference of 33.6 percentage points) also had $P=0.114$. In addition, among the 1,152 Sort=All control cells, ductal and fibroblast cells were both zero, which is inconsistent with the expectation for unsorted whole cells, suggesting dissociation bias; this control is insufficient to represent the composition of healthy lacrimal gland tissue. Fourth, P values from cell-level

differential expression have pseudo-replication and can be used only for candidate ranking. Fifth, both ambient RNA and doublet contamination were present: the doublet detection rate was 9.4%, and doublets were not removed from the main analysis (their impact on the whole-gene differential expression list was not quantified), and ambient RNA could not be fully corrected. Correction was applied only to the case side, biasing the comparison toward “attenuation on the case side”; therefore, no between-group inference was made after correction. Sixth, clinical information was lacking, so no clinical association analysis could be performed. Searches of public databases also did not yield usable same-platform multi-donor controls or ready-made processed matrices. In addition, this study did not perform negative-binomial-model-based simulation validation (comparing the FDR and power of criteria A–D under known ground truth), so the conclusion that “filtering strategy is more critical than filtering stringency” remains observational evidence and awaits quantitative confirmation by simulation studies.

There are three methodological findings. First, whether to take a gene intersection before merging and apply detection-rate filtering was more important than filtering stringency itself in determining functional enrichment conclusions. Without filtering, 53.7% of genes had a detection rate $<5\%$ in at least one group (accounting for 51.1% of significant genes under criterion A), and enrichment was dominated by

cytoskeletal/neural/ion-related terms; after filtering, it shifted to RNA splicing and immune-related terms. The mechanism may be that one-sided zero detection extremizes $\log_2FC$, and such genes are mostly low-expression, tissue-specific transcripts; after filtering removes false differences, widely expressed and stably detected immune-related genes become relatively dominant. This is testable: the one-sided detection rates of top genes under criterion A should be lower than background, and cytoskeletal/neural driver genes should be lower than immune driver genes; however, dominance of immune-related terms does not equal a true signal, because compositional differences alone can produce this effect. Second, threshold-based enrichment and whole-gene ranking GSEA gave inconsistent conclusions, and the immune activation signal from the latter was not reproduced in sample-level analysis, suggesting that it partly came from pseudo-replication and composition differences. Third, ambient RNA caused bidirectional bias, with acinar genes on the control side and B cell markers on the case side. Accordingly, we recommend: report one-sided detection-rate distributions before cross-platform merging; pre-register detection-rate thresholds and perform sensitivity analyses; and perform ambient RNA correction symmetrically on both sides.

The observations in this study both corroborate and supplement existing integration benchmarking studies. Tran et al. [22] systematically compared multiple batch-correction methods and pointed out that method choice has a non-negligible impact on downstream clustering and differential expression; the best practices of Luecken and Theis [11] also emphasize that cross-batch differences should not be directly interpreted after integration. This study does not evaluate integration algorithms themselves but points to an earlier step: whether to take a gene intersection and whether to apply detection-rate filtering before merging may have a greater impact than the choice of subsequent algorithms. Unlike the above studies, which were mostly conducted within “same platform, different batches” or simulated data with credible references, this study faced an extreme situation in which platform and disease were completely confounded and controls were technical replicates; in this situation, integration cannot separate platform effects, so detection-rate filtering becomes the only operable remedy. This observation can be regarded as a

supplement to existing conclusions: the cost of batch effects occurs not only after integration but also during gene-set construction at merging. It should be noted that the technical-gene removal in this study was constructed by symbol-prefix matching and lacks functional specificity—among the 114 genes actually removed under criterion C were three S6 kinases, RPS6KA2, RPS6KA3, and RPS6KB1, as well as MALAT1; therefore, this step should be regarded as conservative noise reduction rather than a functional judgment. Figure 4A and 4B show GSEA within two T cell subsets, CD4⁺ T and CD8⁺ T. They are only methodological demonstrations of within-cell-type analysis and are not used for between-group inference. Within-cell-type comparisons showed that MHC class II molecules in acinar, B, and CD8⁺ T cells showed positive changes on the case side in cell-level analysis. Cell-level indications of T cell receptor signaling components and B cell markers were also observed, but these signals may also come from ambient RNA from immune cells. This section does not propose any mechanistic hypothesis. MHC class II signals in T cells were greatly attenuated after ambient RNA correction, with HLA-DPB1 decreasing from an uncorrected positive value to +0.06 to +0.38, suggesting that this signal may be affected by ambient RNA. Intrinsic T cell activation cannot be inferred. Ligand–receptor analysis was not statistically tested and is only an exploratory description.

Regarding serum miRNA, Nezu et al. [9] reported the top 10 upregulated miRNAs in serum from IgG4-ROD patients versus healthy individuals in the same dataset, GSE155177: miR-550a-5p, miR-642b-3p, miR-4730, miR-4673, miR-4649-5p, miR-4745-5p, miR-128-2-5p, miR-92a-2-5p, miR-3648, and miR-3663-3p. Among the 15 upregulated miRNAs identified here, 10 were identical to these 10, and the $\log_2$ fold changes and raw P values were highly consistent. For example, miR-550a-5p showed 1.54 vs 1.544 and 5×10^{-5} vs 4.45×10^{-5} . This indicates that our analysis reproduced the serum differential profile of the original article. The raw P values of the other five miRNAs, miR-4674, miR-4258, miR-187-5p, miR-3619-3p, and miR-760, were 2.6×10^{-4} to 3.2×10^{-4} , just after the top 10 cutoff in the original article. Their differences mainly came from different probe filtering criteria and thresholds. The original article compared healthy individuals and orbital MALT lymphoma groups and did not report zero-variance

probe handling, so the two studies cannot be directly equated. In this study, all 15 miRNAs were non-significant under the primary criterion excluding zero-variance probes.

Based on the limitations of this study, future work should proceed in the following priority order. First, include same-platform, multi-donor healthy lacrimal gland controls, which is a prerequisite for resolving platform–disease confounding. Second, perform paired lacrimal gland–serum sampling to test the relationship between serum miRNA and local immune changes. Third, add BCR/TCR sequencing and spatial transcriptomics to validate clonal expansion and localize MHC class II-high cells. Fourth, validate key gene expression changes in an independent cohort and include clinical indicators such as serum IgG4, degree of lacrimal gland enlargement, disease duration, and steroid response. In summary, this is an exploratory study with single-donor controls and complete platform–disease confounding. All between-group comparisons were non-significant. The reported differential and enrichment results should be regarded as methodological simulations and hypothesis generation, not reliable biological findings. Methodological conclusions are relatively robust, whereas biological conclusions require validation in independent data with same-platform, multi-donor, paired samples and after contamination correction.

The results of this study cannot be directly used for clinical diagnosis or efficacy assessment: controls were from a single donor, platform and disease were completely confounded, proportion differences were non-significant, and serum miRNAs were non-significant under the primary criterion. The differential diagnosis and treatment decisions for lacrimal gland enlargement should still be based on histopathology and rheumatology guidelines ^[1,23].

4 Conclusions

(1) This study is a methodological evaluation with single-donor controls and complete platform–disease confounding, and it does not report biological conclusions. Because controls were three technical plates from the same donor, the assumption of independent biological replicates was not met. All between-group permutation P values, with the smallest BH-adjusted value of 0.1286, bootstrap intervals, and single-gene sample-level permutations, including acinar HLA-DPB1 $P=0.0556$, the minimum attainable

value for this design but still >0.05 , cannot be used for significance inference and are only descriptive observations.

(2) Methodological conclusion 1: whether to take a gene intersection before cross-platform merging and apply detection-rate filtering is more important than filtering stringency itself in determining enrichment conclusions.

(3) Methodological conclusion 2: ambient RNA can cause bidirectional false signals. In this study, correction was applied only to the case side, so no between-group comparisons were made after correction.

(4) Methodological conclusion 3: the serum miRNA section is only used to reproduce the original workflow and demonstrate methods. Under the all-probe criterion, adjusted $P=0.0476$; under the primary criterion excluding zero-variance probes, values were 0.0651–0.0711, all non-significant. The target gene intersection enrichment fold was only 1.135. This result does not constitute biomarker evidence and does not support any clinical inference.

Author contributions

Zhihao Zhao: Writing–review & editing, Writing–original draft, Visualization, Investigation. Pei Liu: Writing–review & editing, Writing–original draft, Visualization, Investigation. Yu Deng: Writing–review & editing, Investigation. Haibei Dong: Writing–review & editing, Supervision, Funding acquisition, Project administration. Zhiyang Zhang: Writing–review & editing, Supervision, Conceptualization, Funding acquisition, Project administration.

Acknowledgements

This work was supported by the CAPTRA-STAR Scientific Research Fund Project (KY202502041), the Open Fund of Anhui Engineering Laboratory for Conservation and Sustainable Utilization of Traditional Chinese Medicine Resources (TCMRPSU-2022-11), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Conflict of interest

The authors declare that they have no conflict of interest.

References

- [1] Chinese Rheumatology Association. Chinese expert consensus on the diagnosis and treatment of IgG4-related diseases. *Chin J Intern Med.* 2021, 60(3):192-206.
- [2] Bannier-Hélaouët M, Post Y, Korving J, et al. Exploring the human lacrimal gland using organoids and single-cell sequencing. *Cell Stem Cell.* 2021, 28(7):1221-1232.
- [3] Wallace ZS, Naden RP, Chari S, et al. The 2019 American College of Rheumatology/European League Against Rheumatism classification criteria for IgG4-related disease. *Ann Rheum Dis.* 2020, 79(1):77-87.
- [4] Maehara T, Mattoo H, Ohta M, et al. Lesional CD4⁺ IFN- γ ⁺ cytotoxic T lymphocytes in IgG4-related dacryoadenitis and sialoadenitis. *Ann Rheum Dis.* 2017, 76(2):377-385.
- [5] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell.* 2019, 177(7):1888-1902.e21.
- [6] Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol.* 2024, 42(2):293-304.
- [7] Korotkevich G, Sukhov V, Budin N, et al. Fast gene set enrichment analysis. *bioRxiv.* 2021.
- [8] Liu R, Ren T, Xu L, et al. Single-cell atlas of B cell heterogeneity in lacrimal IgG4-related disorders: unraveling the transition from inflammatory exhaustion to clonal malignancy. *FEBS J.* 2026, 293(16):4992-5007.
- [9] Nezu N, Usui Y, Asakage M, et al. Distinctive tissue and serum microRNA profile of IgG4-related ophthalmic disease and MALT lymphoma. *J Clin Med.* 2020, 9(8):2530.
- [10] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods.* 2019, 16(12):1289-1296.
- [11] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol.* 2019, 15(6):e8746.
- [12] Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. *Nat Commun.* 2021, 12(1):5692.
- [13] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol.* 1995, 57(1):289-300.
- [14] Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience.* 2020, 9(12):giaa151.
- [15] Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb).* 2021, 2(3):100141.
- [16] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005, 102(43):15545-15550.
- [17] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015, 43(7):e47.
- [18] Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 2020, 48(D1):D127-D131.
- [19] Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 2019, 8(4):281-291.
- [20] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 2019, 8(4):329-337.
- [21] Yang S, Corbett SE, Koga Y, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* 2020, 21(1):57.
- [22] Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 2020, 21(1):12.
- [23] Perugini CA, Stone JH. IgG4-related disease: an update on pathophysiology and implications for clinical care. *Nat Rev Rheumatol.* 2020, 16(12):702-714.