Simulated single-cell RNA sequencing datasets to inform study design for rare cell type experiments

Cameron Broomfield

The Centre for Bioinformatics and Computational Biology, Stellenbosch University
Email for correspondence: cameronbroomfield17@gmail.com

Supervisor: Dr Elizna Maasdorp

Division of Immunology | School for Data Science and Computational Thinking, Stellenbosch University

Abstract

Background

Single-cell RNA sequencing enables analysis of cellular heterogeneity but remains costly and statistically challenging for rare cell types. Simulators such as scDesign2 can generate synthetic datasets to optimize study design before laboratory investment. This study evaluates scDesign2's ability to reproduce real data characteristics and its usefulness in planning rare cell experiments.

Results

Using a peripheral blood mononuclear cell dataset containing mucosal-associated invariant T cells, simulated datasets were benchmarked against real data for structural, technical and biological fidelity. The simulator reproduced global data topology and gene-level correlations well, but showed weakened cell-cell correlations, particularly in small cell populations. Basic biological signal types such as differential expression and variability were well preserved, while compositional signals were not. A grid search across varying sequencing depth and cell counts revealed that detectability of rare cells scaled irregularly with both parameters, emphasizing trade-offs between cost and performance. Downstream replication of published analyses showed that major transcriptional patterns and immune signatures were retained, though finer scale gene-specific trends were unreliable.

Conclusions

scDesign2 captures broad transcriptomic patterns reliably, but performs less well for rare cell types, where variation and heterogeneity are harder to reproduce. If applied to rare cell studies, its use should be limited to high-level analyses, such as estimating power or assessing overall trends, rather than detailed gene-level interpretation. These results present a practical framework for using simulations to balance sequencing depth, cell number and rare cell detection in future single-cell RNA studies.

Keywords: Singe-cell RNA sequencing; Rare cell detection; Experimental design; Data simulation; MAIT cells; COVID-19

Background

Single-cell RNA sequencing (scRNA-seq) is a relatively new technology, beginning with the whole-transcriptome analysis of a single mouse blastomere cell in 2009 (Tang, Barbacioru, Wang, Nordman, Lee, Xu, Wang, Bodeau, Tuch, Siddiqui, Lao & Surani, 2009). Advancements in high-throughput sequencing and microfluidics saw the number of individual cell transcriptomes sequenced by 2020 approached 60,000 (Jovic, Liang, Zeng, Lin, Xu & Luo, 2022).

scRNA-seq greatly increases the resolution at which we can analyse transcriptomes, revealing differences at the cellular level that would be averaged by the pooled sequencing of bulk RNA sequencing. This single-cell resolution reveals large degrees of cellular heterogeneity, even within the same cell type. It also allows for the identification of rare and transcriptionally distinct populations of cells (Jovic *et al.*, 2022).

One such population is Mucosal-Associated Invariant T (MAIT) cells. MAIT cells are a subset of T cells that express both innate and adaptive immune qualities. MAIT cells make up 1-10% of circulating T cells in the blood, and their levels can fluctuate greatly in response to infection (Nel, Bertrand, Toubal & Lehuen, 2021).

The high dimensionality and technical variability of scRNA-seq data make it challenging to achieve statistical significance, particularly for rare cell types (Zogopoulos, Tsotra, Spandidos, Iconomidou & Michalopoulos, 2025). To addresss this, scRNA-seq simulators are used to generate realistic synthetic data that let researchers plan their experiments *in silico* before investing in participant enrolment, sample collection and laboratory processing (Cao, Yang & Yang, 2021). Typically, these models use gamma, Poisson, Bernoulli and log-normal distributions to reproduce gene counts; however, current simulators still fall short in capturing the complexity of rare populations - motivating this study which evaluates simulator fidelity and its utility in designing rare-cell experiments.

Aims

1) Validate scDesign2 for rare cell analysis
2) Quantify how detection power scales with sequencing depth and cell count
3) Explore the differences between simulated and real datasets in downstream biological analyses

Methods

Data and preprocessing

This project used the peripheral blood mononuclear cell (PBMC) data from Wilk, Rustagi, Zhao, Roque, Martínez-Colón, McKechnie, Ivison, Ranganath, Vergara, Hollis, Simpson, Grant, Subramanian, Rogers & Blish (2020), consisting of 44,721 cells from six Healthy Controls (HC) and seven COVID-19 patients, six of which were admitted to ICU. FASTQ sequences, demultiplexed count matrices and the assembled Seurat object are available for download, with processing scripts available on the authors' GitHub. The data had undergone standard quality control, normalization and cell type annotation as described in the publication. MAIT cells were defined using the marker gene SLC4A10-positive cells (SLC4A10 > 0).

Simulator and parameterization

Simulated datasets were generated using the above single-cell atlas and scDesign2 (Sun, Song, Li & Li, 2021), a statistical simulator that models gene correlations using a Gaussian copula. Cell type proportions were kept identical to the real dataset, while cell count and sequencing depth parameters were varied to evaluate how these experimental design parameters affect data fidelity and the detectability of rare cell types.

Part 1: Real vs Simulated data

Real and simulated datasets, matched by cell count and sequencing depth, were compared to assess the simulator's fidelity in reproducing technical and biological properties in a like-to-like scenario.

Structural Fidelity
Real and simulated datasets were jointly embedded using UMAP after standard data normalisation and scaling. The combined embeddings were used to assess global cluster topology while separate embeddings highlighted MAIT cells to assess their local neighbourhood and overlaps with other cell types.

Quantitative Fidelity
The SimBench *eval_parameter* function quantified how well the simulator reproduced the distributions of technical data characteristics of the real dataset at both the gene level and cell level. At the gene level distributions include mean expression, variance and gene-gene correlation, as well as the mean-variance relationship. At the cell level distributions include effective library size, sparsity and cell-cell correlation, reflecting sequencing depth, dropout rate and heterogeneity.

The distribution of each characteristic was estimated using kernel density estimation (KDE) and the overlap between real and simulated distribution was quantified using integrated squared error (ISE) (Cao *et al.*, 2021). For each metric, the cell types were ranked based on these error scores, and each cell type's metric was plotted against its proportion to assess how cell type abundance influences the property.

Biological Signal Preservation
The SimBench *eval_signal* function quantified how well various biological signal types were retained in the simulation. Each dataset is internally subsetted and each signal type is computed between those subsets. The signals measured include differential expression (DE), differential variability (DV), differential distribution (DD), differential proportion (DP) (Tiberi, Crowell, Samartsidis, Weber & Robinson, 2023) and bimodality (BD). The proportion of genes exhibiting each signal type is compared between datasets, with closer proportions between datasets indicating more faithful biological signal preservation, quantified using symmetric mean absolute percentage error (SMAPE) (Cao *et al.*, 2021).

Gene-Signature Fidelity
Per-cell gene enrichment scores were computed using UCell and a curated 167-gene MAIT cell gene-signature (Garner, Amini, FitzPatrick, Lett, Hess, Filipowicz Sinnreich, Provine & Klenerman, 2023), and score distributions were compared between real and simulated datasets to assess the preservation of gene-level signature patterns.

Part 2: Grid search

A grid search over the cell count and sequencing depth parameters was performed to investigate how design parameters influence data fidelity and rare cell detectability. The simulated cell counts and sequencing depths both ranged from 0.5x-2x the original value of the real dataset, with step sizes of 0.25x for a total of 49 simulation steps.

Each simulated dataset was assessed for how distinctly the rare cell type stands out among other cell types. Every cell was scored against the 167-gene MAIT signature from Garner et al. (2023) and ranked accordingly. The highest ranked fraction ('K', corresponding to the expect MAIT cell fraction) of cells were compared to ground-truth labels provided by the simulator and quantified for precision, recall, F1 and enrichment for MAIT cells, assessing how well the top-K cells captured the true MAIT cell population. This approach is popular in applications with severe class imbalance (Kar, Narasimhan & Jain, 2015)(Paton, Boiko, Perkins, Cemalovic, Reschützegger, Gomes & Narayan, 2025).

This optimization-style framework assesses how well MAIT cells are represented amongst the highest scoring fraction of cells, and investigates if certain combinations of cell-count and sequencing depth allow us to better study them.

Part 3: Application/Replication

To assess how well the simulated datasets support downstream biological tasks, a set of analyses were adapted from published MAIT cell scRNA studies. The real dataset and two simulated datasets at different cell counts and sequencing depth were chosen based on the outcome of the grid search. This section explored how variations in experimental design parameters affect those outcomes.

Functional immune signatures: gene panels representing cytotoxicity, regulatory metabolism, interferon response, exhaustion, S100 gene family and apoptosis signatures were scored using UCell and visualised to assess signature patterns (Qi, Zhang, Huang, Fu & Zhao, 2021).

Differential gene expression: comparisons were performed on MAIT cells between groups (Healthy Control vs Severe COVID-19) using the FindMarkers function with the MAST algorithm in Seurat. Genes were considered significantly upregulated if the logarithm of the fold-change (logFC) was >0.25 and adjusted p value <0.01 (Yang, Wen, Qi, Gao, Chen, Xu, Wei, Wang, Tang, Lin, Zhao, Zhang, Zhang & Zhang, 2021).

Gene panel: for select genes, mean expression was compared across healthy and severe COVID-19 groups across the real and simulated datasets (Yang *et al.*, 2021).

Computing

All analyses were conducted in R v4.5.1 (R Core Team, 2021), using Seurat v5.3.0 (Hao, Stuart, Kowalski, Choudhary, Hoffman, Hartman, Srivastava, Molla, Madad, Fernandez-Granda & Satija, 2024) for preprocessing, clustering and differential expression analysis, scDesign2 v0.1.0 (Sun *et al.*, 2021) for data simulation, SimBench v0.99.1 (Cao *et al.*, 2021) for fidelity and signal evaluation, and UCell v2.13.1 (Andreatta & Carmona, 2021) for gene signature scoring.

Results

R1. Real vs Sim

The real and simulated dataset, matched in cell count (44,721 cells) and sequencing depth, were jointly embedded to visualize the overlap between the two datasets (**Fig. 1**). The overall topology is preserved between the real and the simulated dataset, with the same number and arrangement of both major and minor clusters. The simulated cells cluster more tightly around their centroids, with sharper cluster boundaries compared to the real cells. As a result, the simulated points fall largely within the boundaries of the corresponding real clusters.

The shape of clusters is also well retained, most notably the agreement between elongated clusters in the real and simulated data. The real data has transitions between some of its clusters, reflecting a continuum of transcription. These continua are not well preserved in the simulated data.

Overall, the simulation reproduces global and local structure accurately, although some local properties are lost. With the global structure validated, the next analysis focuses on MAIT cells and their behaviour in this framework.

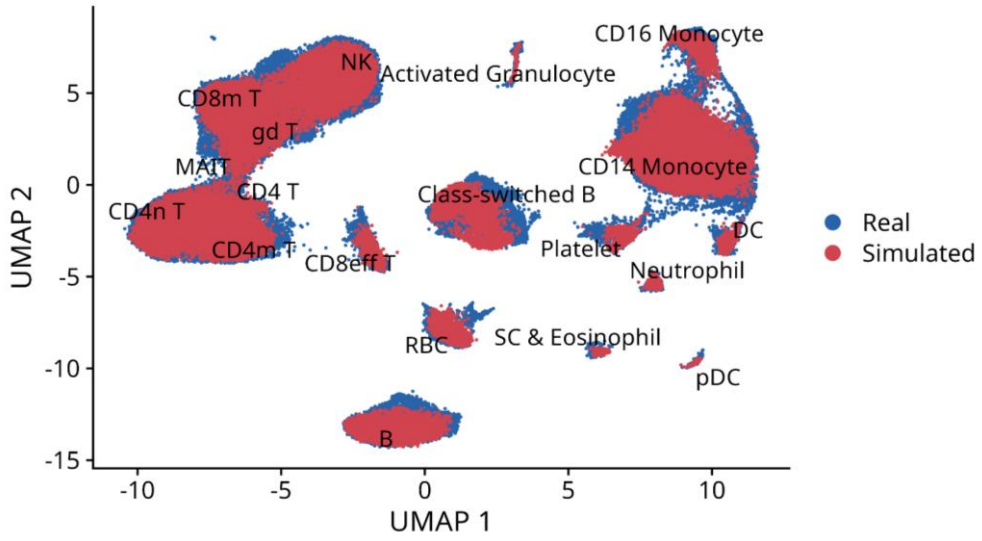

Figure 1: Joint UMAP embedding of real and simulated data under matched cell counts and sequencing depth. Real cells are coloured red, simulated cells are blue. Clusters are labelled by their cell type.

Each dataset was then embedded separately into its own UMAP plot, and MAIT cells (294 cells) were highlighted to visualize their localization (**Fig. 2**). In the real dataset, MAIT cells localise close to γδ T cells, spanning the boundaries of CD8 and CD4 T cell clusters. A substantial proportion of MAIT cells are interspersed among CD4 T cells, with a smaller portion occurring within the CD8 effector cluster, and an even smaller set among CD8 memory T and NK cells.

In the simulated dataset, the MAIT cells form two major clusters – one bordering the CD8m T cell cluster and another interspersed within the C4m T cell cluster. Notably, the distinct MAIT cell subpopulation within the CD8 effector cluster observed in the real dataset is absent in the simulation.

These findings suggest that while the simulation maintains the relative proximity of major T cell clusters, a large portion of the heterogeneity within the MAIT population is lost. The absence of the set within CD8 effector T cells underscores the possibility of losing smaller but meaningful groups.

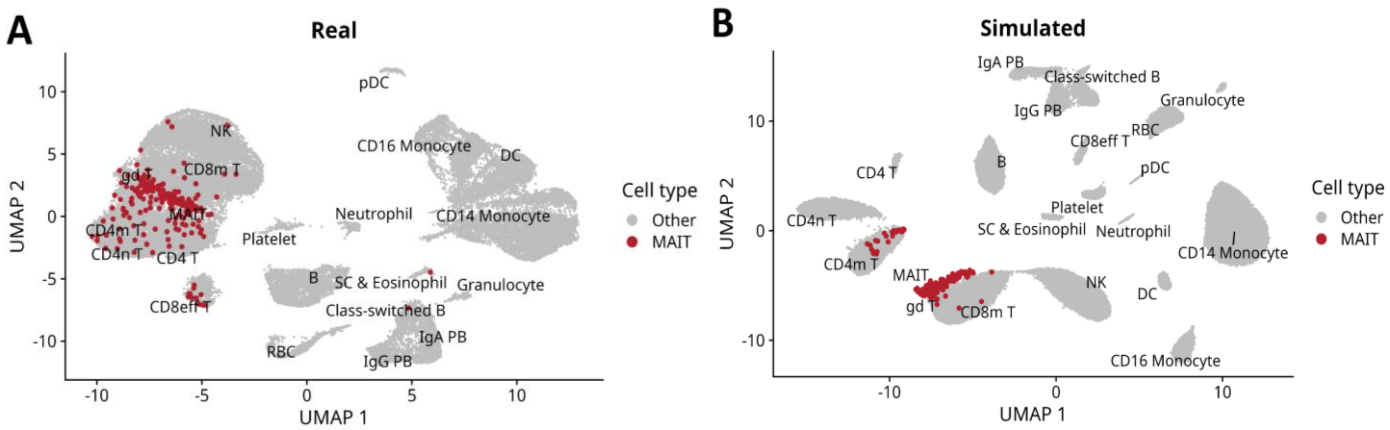

Figure 2: Separate UMAP embeddings of (**a**) Real and (**b**) Simulated cells. MAIT cells are coloured red. Clusters are labelled with their cell type.

To build upon these observations, quantitative fidelity measurements, using SimBench's *eval_parameter* function, were applied to **MAIT cells only** (**Fig. 3**).

Cell types were ranked by their agreement with the corresponding real distributions. Although their performance varied across properties, MAIT cells generally ranked in the midrange and often exceeded several more abundant cell types' performance, with a combined fidelity ranking of **11 of 21** cell types.

At the gene level, the simulation performed well - gene-wise mean, variance and gene-gene correlation (**Fig 3a, b and c respectively)** were effectively preserved and the mean-variance trend showed the expected negative binomial pattern.

At the cell level, the effective library size of MAIT cells was accurately reproduced, whereas, per-cell sparsity (**Fig. 3d, e**) was poorly replicated, a pattern that improves with increasing cell type proportion. The weakest metric for MAIT cells was the cell-cell correlation (**Fig. 3f**), indicating limited preservation of cellular heterogeneity within the MAIT cell type, likely due to their small sample size.

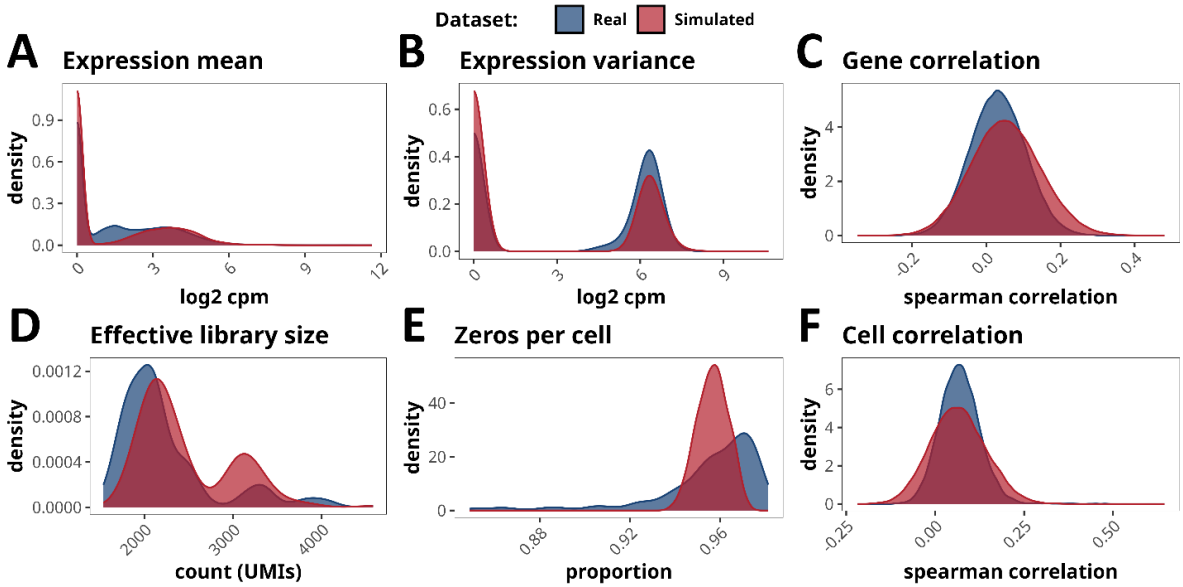

Figure 3: Kernel-density estimated distributions of fidelity metrics in real and simulated MAIT cells. Metrics include (**A**) mean gene expression, (**B**) variance of gene expression, (**C**) spearman correlation between genes, (**D**) effective library size, (**E**) proportion of zeros per cell and (**F**) spearman correlation between cells.

To evaluate how well the simulation reproduces **global** biological signals observed in the real dataset, SimBench's *eval_signal* function was applied (**Fig. 4**).

The proportion of DE, DV, DD and BD genes in the real and simulated datasets aligned closely, with SMAPE scores 0.995 and above for these signals. DP is the only signal that differs substantially, with a SMAPE score of 0.851.

Overall, these results show that the simulation can replicate basic expression shifts in mean (DE) and variation (DV), as well as more complex distributions measured by the DD and BD signals but struggles with the compositional signal DP, which captures higher order gene- and cell-level phenomena.

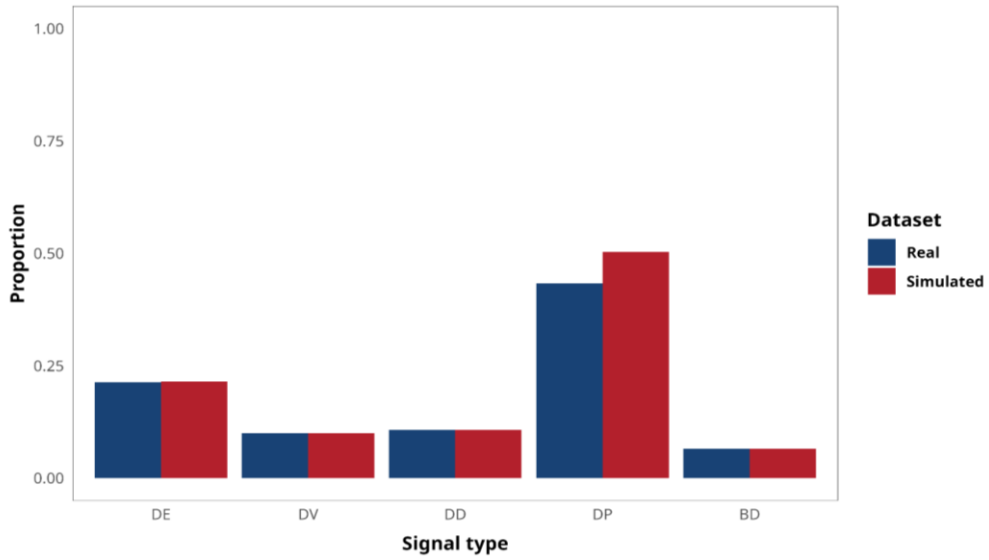

Figure 4: Comparison of biological signal types between real and simulated datasets. Proportions include differential expression (DE), differential variability (DV), differential distribution (DD), differential proportion (DP), and bimodality (BD).

With the simulator's ability to replicate major categories of biological signal, we then investigated cell type-specific signals in gene expression signatures.

UCell was used to compute per-cell gene set enrichment scores using a curated MAIT cell gene signature (Garner *et al.*, 2023). The distribution of these scores is summarized by cell type for both the real and simulated datasets (**Fig. 5**).

In the real dataset, MAIT cells display the most distinct enrichment pattern, with some overlap with CD4m, CD4n and γδ T cells due to shared transcriptional features. In the simulated dataset, this overall structure is well preserved - MAIT cells remain the strongest signal - though the distributions are tighter and shifted down, indicating reduced variability but retention of key enrichment patterns.

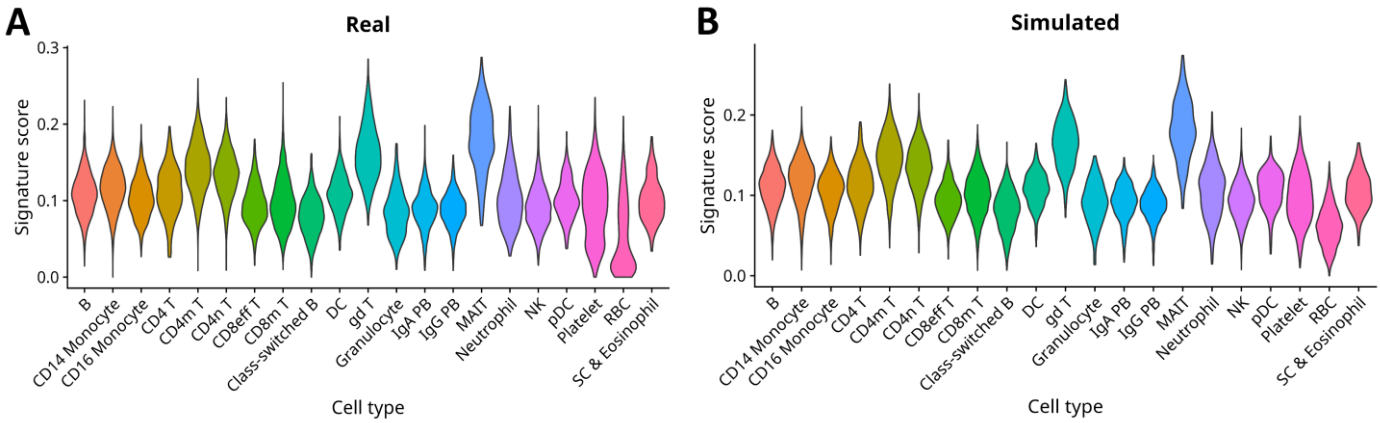

Figure 5: MAIT gene signature score for each cell types in (**A**) real and (**B**) simulated datasets showing elevated distribution of scores in MAIT cell population retained by the simulator.

R2. Grid Search

A grid of simulations spanning different sequencing depths and cell counts was generated, and each simulation was evaluated for data fidelity and rare-cell detectability. Fidelity metrics such as the proportion of MAIT cells and percentage of mitochondrial genes were consistent across all simulations, indicating stable simulator behaviour across the grid.

MAIT cell detectability was quantified by the curated 167-gene signature score of each simulation step. This score simply scaled up with sequencing depth and inversely with cell count (**Fig 2a**). Instead, enrichment@K was quantified to assess how well the top ranked fraction of cells cover the true MAIT population (**Fig 2b**).

The enrichment and F1 heatmaps did not show a smooth gradient over the grid, showing that detection performance scaled irregularly with sequencing depth and cell count. Because both parameters affect trade-offs in cost and sampling coverage, the relative differences between grid points are more meaningful than absolute values. Higher scores at low cell counts or high sequencing depths sacrifice coverage and cost respectively. Based on these relative comparisons, two representative simulations were selected for further analysis in the next step.

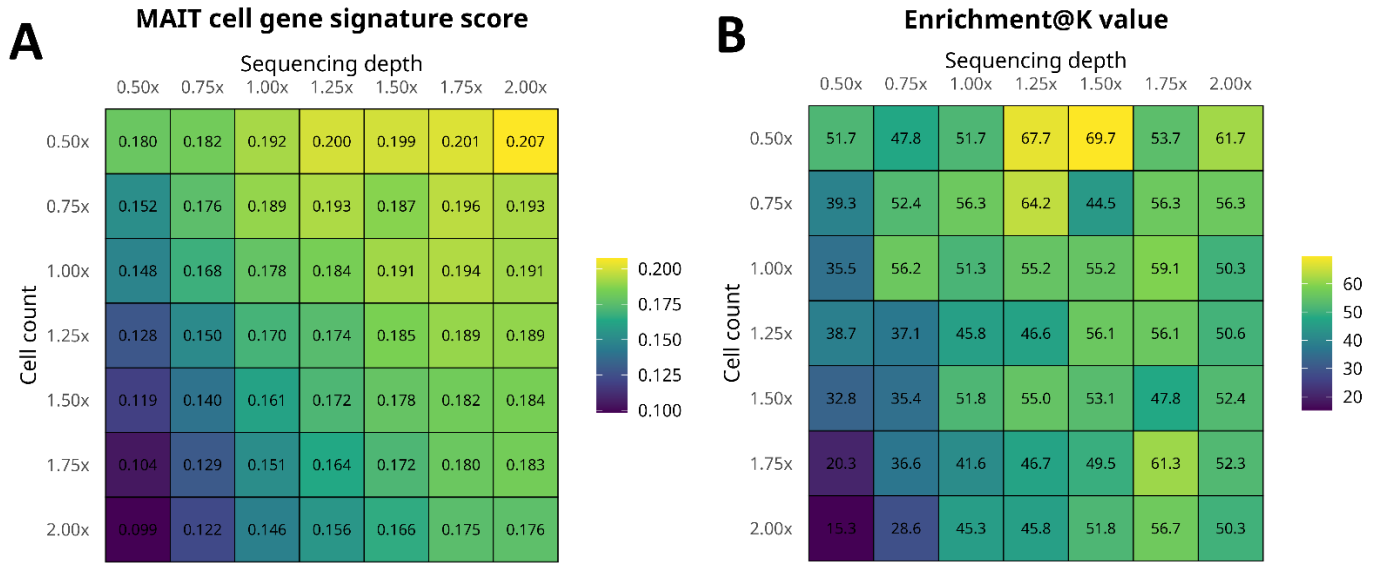

Figure 6: Heatmaps showing (**A**) MAIT cell gene signature score and (**B**) enrichment@K at varying sequencing depth and cell count.

R3. Application/Replication

This section explores how simulated datasets perform in downstream biological analyses compared to the real data. Two simulations were selected from the grid search - differing in cell count and sequencing depth (1.5x cell, 1x and 2x sequencing depth) but showing similar F1 and enrichment scores - to examine how experimental design choices affect downstream interpretation when overall detection performance is comparable.

Functional immune signatures

Across the real dataset and both simulated datasets, the MAIT cell population clustered into three communities, and the immune cell functional gene signature patterns are visualised in **Figure 7**.

The most remarkable pattern seen in the functional signatures is the presence of a cytotoxic-gene enriched, and a regulatory metabolism-gene enriched cluster with little overlap between the two. This pattern, as well as the small S100-gene enriched cluster is preserved in the simulated datasets.

Notable, however, is the loss of complexity in the second simulated dataset. The clusters therein are much smoother and more separated. This degradation is not present in the first simulation, which suggests that the simulator struggles more with simulating increased sequencing depth than increased cell counts.

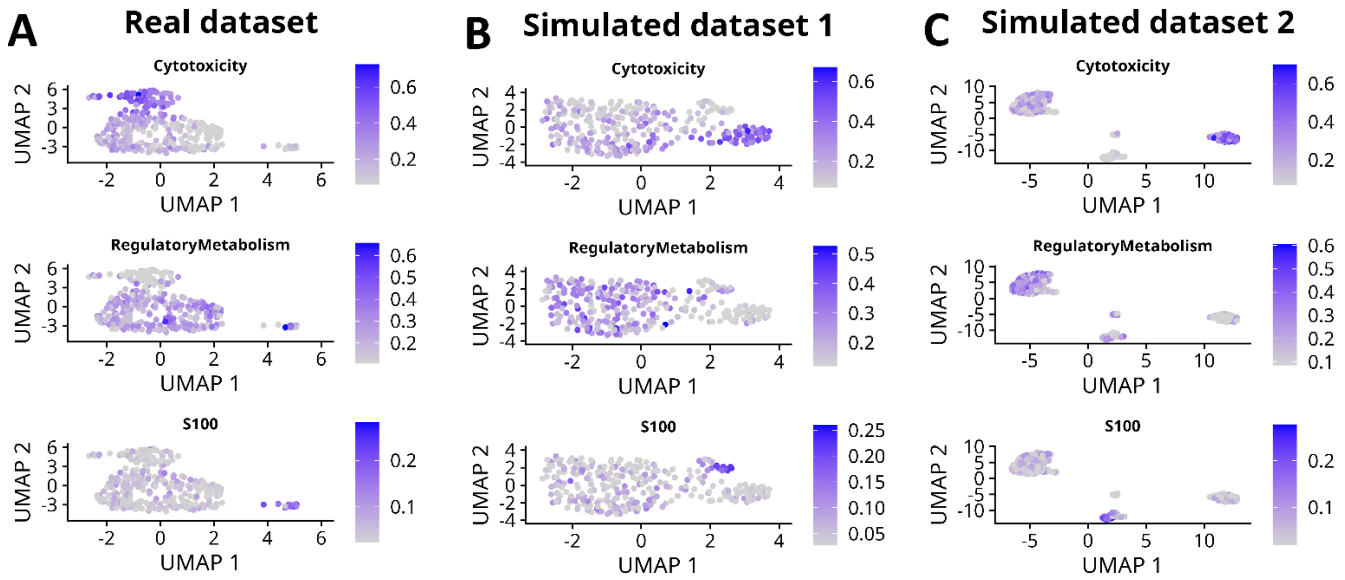

Figure 7: Replication of MAIT cell immune functional signatures in the (**A**) real dataset, the (**B**) first simulated dataset at 1.5x cell count and 1x sequencing depth, and the (**C**) second simulated dataset 1.5x cell count and 2x sequencing depth.

Differential gene expression

Differential gene expression analysis was performed on MAIT cells between healthy and severe COVID-19 patients and visualised in volcano plots (**Fig. 8**).

The broad pattern of genes was consistent across the real and simulated datasets. The real dataset had nineteen significant genes, while the simulations had seventeen and 54 respectively (thirteen common across datasets), following the increasing sequencing depth trend. The log fold-changes were similar across datasets, indicating preservation of relative shifts in expression. The adjusted p-values increased in significance with sequencing depth. Overall, the simulator maintained general gene distribution and relative trends but showed some propensity to miss certain genes.

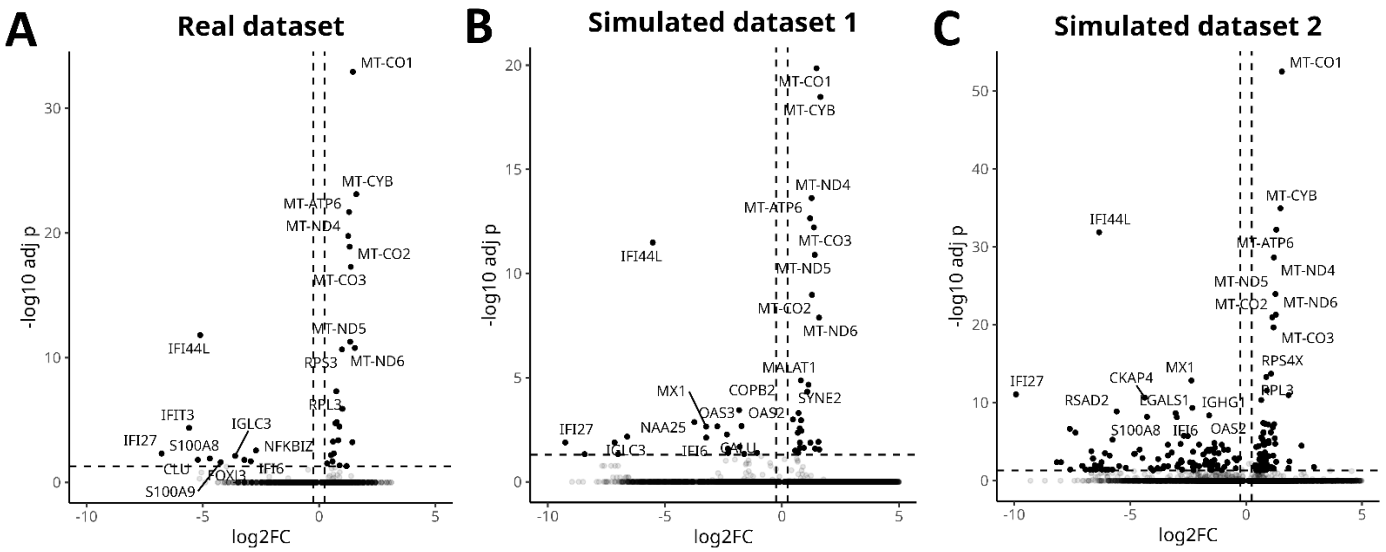

Figure 8: Comparison of differential gene expression in MAIT cells between Healthy Controls and Severe COVID-19 patients shown as volcano plots of the (**A**) real dataset, (**B**) simulated dataset 1, and (**C**) simulated dataset 2.

Gene panel

Mean expression patterns of selected genes were compared between healthy and severe COVID-19 MAIT cells across the datasets (**Fig. 9**). Overall, preservation of gene-level trends was mixed, reflecting biological variability and simulation uncertainty. IFITM3 and S100A8 showed a consistent pattern across datasets. DUSP2 and JUND had moderate agreement, while the pattern exhibited between groups by the RORC gene was inverted across the simulated datasets. These shows that the simulator's performance on the individual gene level is mixed and has the potential to diverge substantially.

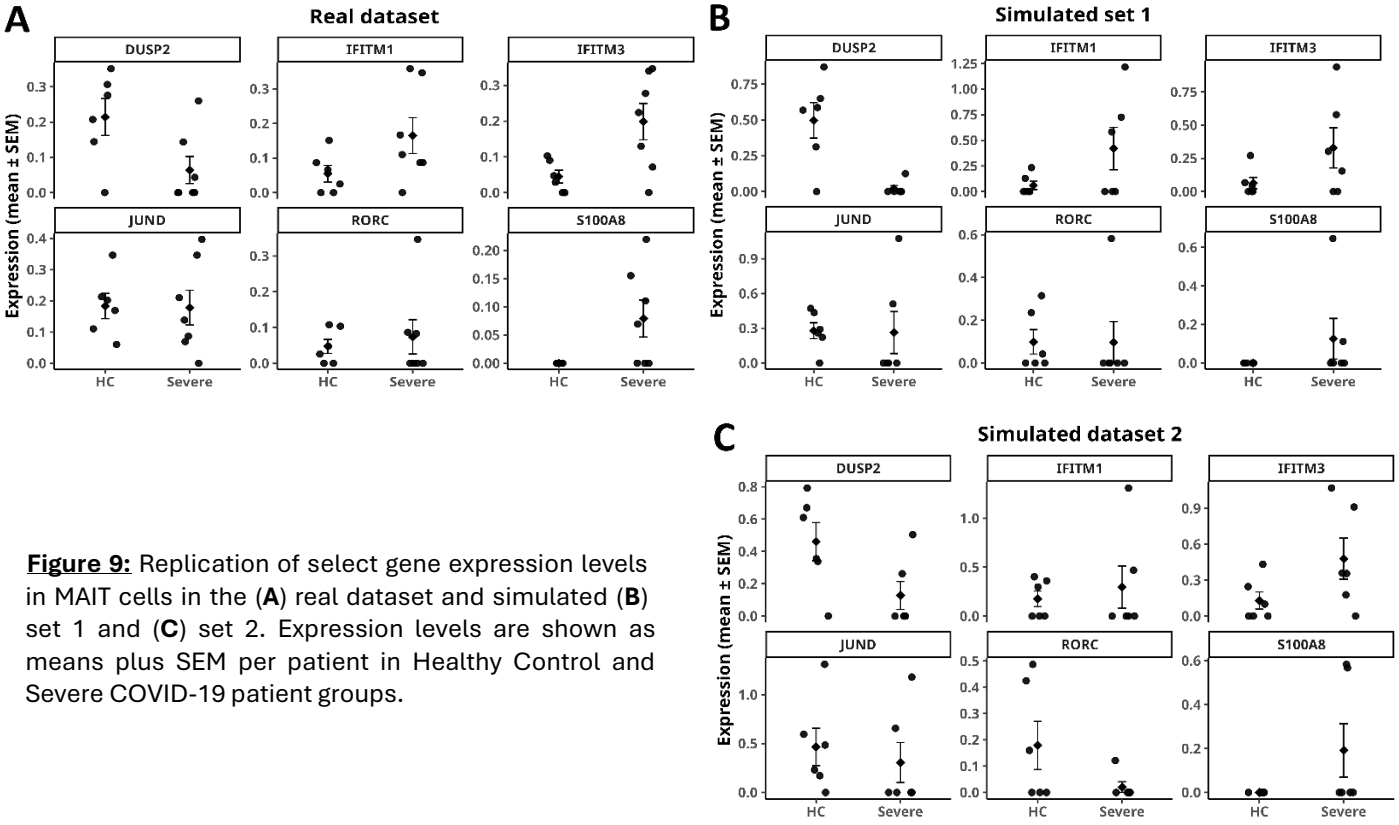

Figure 9: Replication of select gene expression levels in MAIT cells in the (**A**) real dataset and simulated (**B**) set 1 and (**C**) set 2. Expression levels are shown as means plus SEM per patient in Healthy Control and Severe COVID-19 patient groups.

Discussion

Part 1 benchmarks the simulator against a real single-cell dataset under matched conditions to validate whether it can accurately reproduce the real dataset's data fidelity and biological signals.

At the global level, the preserved topology of clusters seen in the joint UMAP embedding confirm that scDesign2 retains broad relationships between cell types - a finding expected because it uses explicit cell type labels when modelling. Continuous trajectories between cell types are absent in the simulated data, indicating limited ability for scDesign2 to capture the differentiation process, though this task is generally left to specialized cell differentiation models (Cannoodt, Saelens, Deconinck & Saeys, 2021) (Papadopoulos, Gonzalo & Söding, 2019).

The loss of biological complexity is illustrated by the simple and distinct clusters in the simulated data. Promisingly, the simulator still captured some of the heterogeneity observed within the MAIT cell population; however, the loss of the CD8 effector-like MAIT subset is significant, given that they are a significant population with increased pro-inflammatory and cytotoxic capacity (Fang et al., 2025).

The simulator reproduced library size distributions well, however sparsity patterns were less realistic. While the average fraction of zeros per cell matched, the underlying distributional shapes differed. This is notable because scDesign2 selects the best of four distributions to model gene dispersion (Poisson, Zero-Inflated Poisson, Negative Binomial and Zero-Inflated Negative Binomial) individually for each gene (Sun *et al.*, 2021). Even with this data-driven model selection, scDesign2 still fails to reproduce realistic dropout patterns.

The simulator showed strong performance in its gene-gene correlation, due to its use of a Gaussian copula for modelling gene dependencies. In contrast, cell-cell correlation was poorly reproduced in MAIT cells, a function of their small population size and indicating limited fidelity in rare cell population structure.

The biological signal evaluation showed that the simulator can model major categories of single-cell biological variation at a global level. Consistent with this, the MAIT gene signature analysis showed the cell type specific feature patterns are largely retained through simulation.

Overall, these findings are consistent with scDesign2's design, which, unlike earlier simulators, combines gene preservation and gene correlation while permitting variation in sequencing depth and cell number (Sun *et al.*, 2021).

The grid search in **Part 2** was designed as a practical tool to 1) validate the simulator's performance over a range of parameters and 2) to enable researchers to assess how sequencing depth and cell count affect rare cell detectability and thus better plan for their experiments. The notebook containing the grid search, along with all other analyses in this report can be accessed via GitHub linked in the "Availability of data and materials" section.

The set of fidelity metrics remained stable across all simulation steps, either remaining constant, or scaling in expected ways with the parameters, showing the simulator's robust performance under different scenarios.

The rare cell detectability, however, scaled irregularly (non-smoothly), suggesting complex distributional shifts underlying the changing parameters of cell count and depth. Increasing the number of cells may hurt rare cell detectability, aligning with previous work studying this trade off (Svensson, da Veiga Beltrame & Pachter, 2019). Furthermore, while detectability generally increased with higher sequencing depth, in some scenarios this started to show negative returns.

It is the relative differences between these detectability metrics that is important, thus the grid search serves to balance trade-offs, rather than optimize a single solution. The heatmaps offer a way to help researchers balance performance, cost and scope, aiming to guide experimental planning using empirical analyses rather than assumptions, although their applicability would benefit validation on additional real-world datasets.

To explore this further, and to demonstrate their application to downstream analyses, two simulations with similar detectability values were analysed in part 3.

Part 3 of this study took two simulations that performed similarly in detectability metrics but differ significantly in input parameters (1.5x cell, 1x and 2x sequencing depth) and aims to explore the utility of simulation based experimental design, beyond the fidelity of the simulated data.

The first step examined whether these parameter differences altered the transcriptional landscape of MAIT cells, using immune cell functional gene signature feature plots to visualise key immune functions. Both simulations reproduced the broad structure seen in the real dataset, notably the separation between the cytotoxic and regulatory-metabolic subpopulations. This retention of heterogeneity is significant, as MAIT cell transcriptomes are known to diverge once reaching maturity (Chandra, Ascui, Riffelmacher, Chawla, Ramírez-Suástegui, Castelan, Seumois, Simon, Murray, Seo, Premlal, Schmiedel, Verstichel, Li, Lin, Greenbaum, Lamberti, Murthy, Nigro, Cheroutre, Ottensmeier, Hedrick, Lu, Vijayanand & Kronenberg, 2023).

What is striking however, is the complete loss of continuum between these subpopulations within the simulated dataset at 2x sequencing depth. This is a characteristic of reference-based simulators that are limited to the complexity of the input dataset and thus struggle to extrapolate beyond observed conditions (Crowell, Morillo Leonardo, Soneson & Robinson, 2023).

The differential expression analysis yielded promising results. Doubling the sequencing depth expectedly increased the number of genes above the significance threshold and the degree of significance in already

identified genes (Rasim Barutcu, 2024). Noteworthy, however, the logFC values remained stable - consistent with expectations, since sequencing depth should not influence biological effect size (Love, Huber & Anders, 2014). This stability, emerging despite logFC not being explicitly modelled, highlights the simulators robustness in this higher-order property.

Assessment at the gene level indicated that while global transcriptional patterns are well preserved - largely in part due to the Gaussian copula's correlation modelling - the reliability of individual gene-level behaviour is limited. These limitations align with observations in the field that there is still no consensus on the most appropriate statistical distribution to model gene expression in scRNA-seq data (Duo, Li, Lan, Tao, Yang, Xiao, Sun, Li, Nie, Zhang, Liang, Liu, Hao & Li, 2024). While hybrid strategies have been proposed and, in scDesign2's case, implemented, this remains an area for improvement.

Conclusions

The simulator achieved high overall fidelity, reproducing global data structure and major categories of biological signal, including within the rare cell type. Gene-level properties were well modelled, but sparsity patterns and cell-cell correlations were less faithfully captured, likely reflecting the simulator's limited ability to handle variability in small cell populations. While some higher level summary features of the MAIT cells were unexpectedly well preserved, the fidelity of individual features remained weak, with inconsistent gene-level variability patterns limiting confidence.

These results indicate that scDesign2 can reliably emulate broad trends and averages but lacks realism at finer resolutions and more variable biological conditions. Consequently, while it shows potential for power estimation decisions, single-cell RNA-seq simulation should be used for high-level analyses rather than specific gene-level interpretations in rare cells.

Availability of data and materials

The COVID-19 single-cell atlas dataset used in this study was original published by Wilk et al. (2020) with code available at the authors' GitHub repository (https://github.com/ajwilk/2020_Wilk_COVID), with raw sequencing data available at NCBI Gene Expression Omnibus (accession no. GSE150728).

Code for simulations and analyses done in this paper are available on GitHub (https://github.com/camfam17/Single-Cell-RNA-Simulation).

References

1) Andreatta, M. & Carmona, S.J. 2021. UCell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal*. 19:3796–3798. DOI: 10.1016/j.csbj.2021.06.043.

2) Cannoodt, R., Saelens, W., Deconinck, L. & Saeys, Y. 2021. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*. 12. DOI: 10.1038/s41467-021-24152-2.

3) Cao, Y., Yang, P. & Yang, J.Y.H. 2021. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nature communications*. 12(1):6911. DOI: 10.1038/s41467-021-27130-w.

4) Chandra, S., Ascui, G., Riffelmacher, T., Chawla, A., Ramírez-Suástegui, C., Castelan, V.C., Seumois, G., Simon, H., Murray, M.P., Seo, G.Y., Premlal, A.L.R., Schmiedel, B., Verstichel, G., Li, Y., Lin, C.H., Greenbaum, J., Lamberti, J., Murthy, R., Nigro, J., Cheroutre, H., Ottensmeier, C.H., Hedrick, S.M., Lu, L.F., Vijayanand, P. & Kronenberg, M. 2023. Transcriptomes and metabolism define mouse and human MAIT cell populations. *Science Immunology*. 8. DOI: 10.1126/sciimmunol.abn8531.

5) Crowell, H.L., Morillo Leonardo, S.X., Soneson, C. & Robinson, M.D. 2023. The shaky foundations of simulating single-cell RNA sequencing data. *Genome Biology*. 24(1):62. DOI: 10.1186/s13059-023-02904-1.

6) Duo, H., Li, Y., Lan, Y., Tao, J., Yang, Q., Xiao, Y., Sun, J., Li, L., Nie, X., Zhang, X., Liang, G., Liu, M., Hao, Y. & Li, B. 2024. Systematic evaluation with practical guidelines for single-cell and spatially resolved transcriptomics data simulation under multiple scenarios. *Genome Biology*. 25. DOI: 10.1186/s13059-024-03290-y.

7) Fang, Y., Chen, Y., Niu, S., Lyu, Z., Tian, Y., Shen, X., Li, Y.R. & Yang, L. 2025. Biological functions and therapeutic applications of human mucosal-associated invariant T cells. BioMed Central Ltd. DOI: 10.1186/s12929-025-01125-x.

8) Garner, L.C., Amini, A., FitzPatrick, M.E.B., Lett, M.J., Hess, G.F., Filipowicz Sinnreich, M., Provine, N.M. & Klenerman, P. 2023. Single-cell analysis of human MAIT cell transcriptional, functional and clonal diversity. *Nature Immunology*. 24:1565–1578. DOI: 10.1038/s41590-023-01575-1.

9) Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C. & Satija, R. 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Research. DOI: 10.1038/s41587-023-01767-y.

10) Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. & Luo, Y. 2022. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and translational medicine*. 12:e694. DOI: 10.1002/ctm2.694.

11) Kar, P., Narasimhan, H. & Jain, P. 2015. Surrogate functions for maximizing precision at the top, in *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, International Machine Learning Society (IMLS). 189–198. Available: https://arxiv.org/abs/1505.06813 [2025, November 02].

12) Love, M.I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 15. DOI: 10.1186/s13059-014-0550-8.

13) Nel, I., Bertrand, L., Toubal, A. & Lehuen, A. 2021. MAIT cells, guardians of skin and mucosa? Springer Nature. DOI: 10.1038/s41385-021-00391-w.

14) Papadopoulos, N., Gonzalo, P.R. & Söding, J. 2019. PROSSTT: Probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics*. 35:3517–3519. DOI: 10.1093/bioinformatics/btz078.

15) Paton, A.E., Boiko, D.A., Perkins, J.C., Cemalovic, N.I., Reschützegger, T., Gomes, G. & Narayan, A.R.H. 2025. Connecting chemical and protein sequence space to predict biocatalytic reactions. *Nature*. 646:108–116. DOI: 10.1038/s41586-025-09519-5.

16) Qi, F., Zhang, W., Huang, J., Fu, L. & Zhao, J. 2021. Single-Cell RNA Sequencing Analysis of the Immunometabolic Rewiring and Immunopathogenesis of Coronavirus Disease 2019. *Frontiers in Immunology*. 12. DOI: 10.3389/fimmu.2021.651656.

17) R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available: https://www.R-project.org/ [2025, November 01].

18) Rasim Barutcu, A. 2024. Evaluation of Replicate Number and Sequencing Depth in Toxicology Dose-Response RNA-seq. *Computational Toxicology*. 30. DOI: 10.1016/j.comtox.2024.100307.

19) Sun, T., Song, D., Li, W.V. & Li, J.J. 2021. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*. 22(1):163. DOI: 10.1186/s13059-021-02367-2.

20) Svensson, V., da Veiga Beltrame, E. & Pachter, L. 2019. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. DOI: 10.1101/762773.

21) Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. & Surani, M.A. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*. 6:377–382. DOI: 10.1038/nmeth.1315.

22) Tiberi, S., Crowell, H.L., Samartsidis, P., Weber, L.M. & Robinson, M.D. 2023. distinct: A NOVEL APPROACH TO DIFFERENTIAL DISTRIBUTION ANALYSES. *Annals of Applied Statistics*. 17:1681–1700. DOI: 10.1214/22-AOAS1689.

23) Wilk, A.J., Rustagi, A., Zhao, N.Q., Roque, J., Martínez-Colón, G.J., McKechnie, J.L., Ivison, G.T., Ranganath, T., Vergara, R., Hollis, T., Simpson, L.J., Grant, P., Subramanian, A., Rogers, A.J. & Blish, C.A. 2020. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine*. 26:1070–1076. DOI: 10.1038/s41591-020-0944-y.

24) Yang, Q., Wen, Y., Qi, F., Gao, X., Chen, W., Xu, G., Wei, C., Wang, H., Tang, X., Lin, J., Zhao, J., Zhang, M., Zhang, S. & Zhang, Z. 2021. Suppressive Monocytes Impair MAIT Cells Response via IL-10 in Patients with Severe COVID-19. *The Journal of Immunology*. 207:1848–1856. DOI: 10.4049/jimmunol.2100228.

25) Zogopoulos, V., Tsotra, I., Spandidos, D., Iconomidou, V. & Michalopoulos, I. 2025. Single-cell RNA sequencing data dimensionality reduction (Review). *World Academy of Sciences Journal*. 7(2):27. DOI: 10.3892/wasj.2025.315.