Comparing single cell sequencing data of Zika, Influenza viruses, and SARS-CoV-2 employing semantic categorizations

F. D. **Keogh**, A. **Dietz**, R. **Koenig**
Institute of Infectious Diseases and Infection Control, Jena University Hospital, Germany
Correspondance: rainer.koenig@uni-jena.de

BACKGROUND

Viruses extensively reprogram gene exprssion in host cells to support their replication and persistence. Identifying conserved gene regulatory host responses across different viral infections is therefore essential for understanding pathogenic mechanisms and for discovering potential host-directed therapeutic targets. Recent studies (Reffsin *et al.* 2025) show that viral susceptibility can vary strongly between individual cells, even within the same cell type. Intrinsic transcriptional cell states can determine whether a cell becomes infected or remains resistant, highlighting substantial heterogeneity in host responses at the single-cell level. Advances in single-cell RNA sequencing enable detailed analysis of virus-induced transcriptional changes at cellular resolution. However, comparing functional responses across datasets and pathogens remains challenging because transcriptomic studies typically generate large lists of enriched gene sets.. To address this, we applied **Slimformer**, an embedding-based natural language processing framework for semantic gene set categorization. Slimformer integrates gene set names, descriptions, hierarchical relations, and associated genes from Gene Ontology to generate semantic embeddings and assigns gene sets to broader biological categories tailored for virology. Using this approach, we analyzed scRNA-seq datasets from human cells infected with **Zika virus, Influenza A virus,** and **SARS-CoV-2**, enabling a systematic comparison of enriched biological processes and host responses across diverse viral infections.

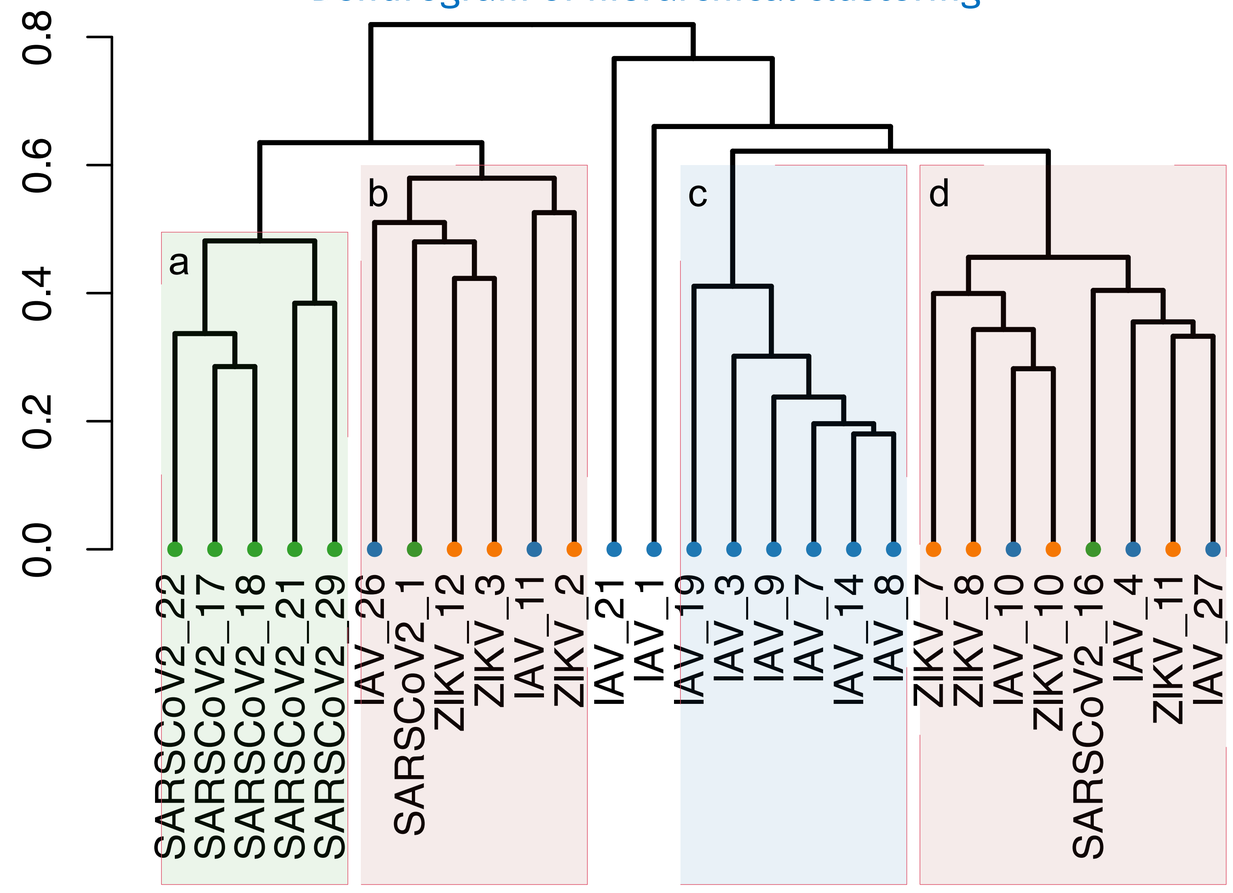
METHODS/RESULTS

1. **Integration**
   scRNA-seq datasets were analysed with scanpy and integrated per virus using scVI.
2. **Differential Expression**
   Infected vs. mock clusters were compared using the Wilcoxon test.
3. **Gene Set Enrichment**
   Enrichment was performed with fgsea, clusterProfiler, and g:Profiler.
   Gene sets supported by ≥2 tools were retained.
4. **Clustering**
   Gene sets were categorized with **Slimformer**. A weighted gene set x cluster distance matrix was computed and hierarchically clustered to identify shared host responses.

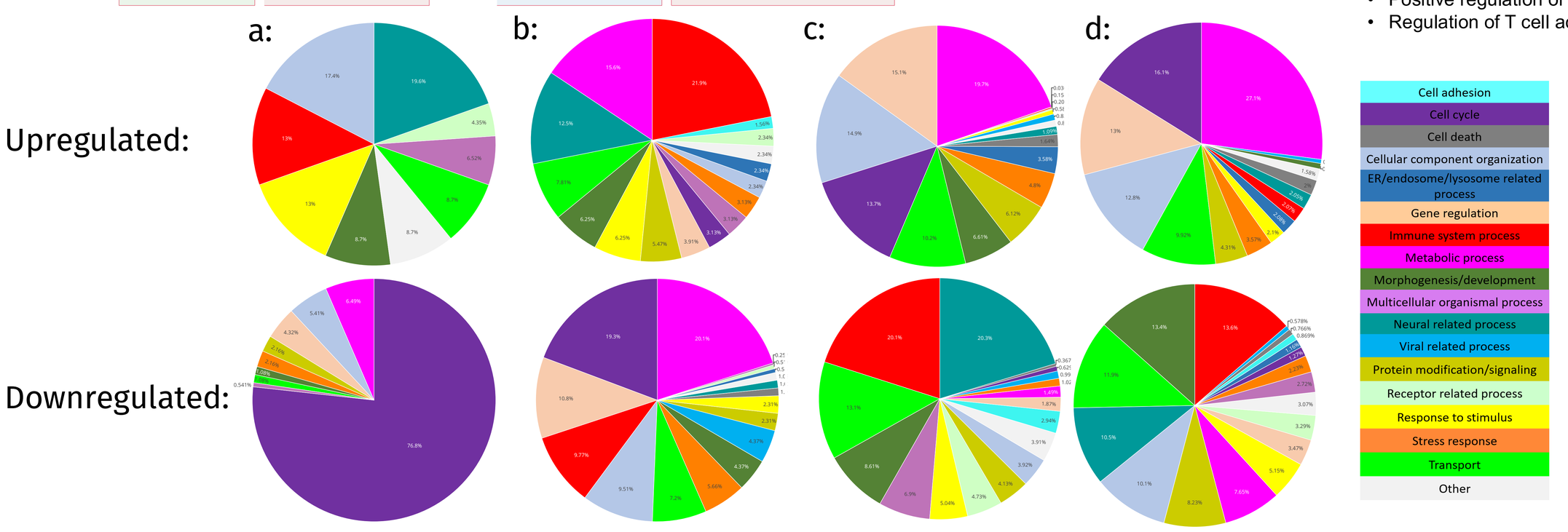
Used publicly available transcription profiles

	Ravindra et al., 2021	Nilsson-Payant et al., 2021	Fiege *et al.*, 2019	Karim et al., 2025	Roach *et al.*, 2024	Russell *et al.*, 2018	Wang *et al.*, 2020	Rivera-Cardona *et al.*, 2026	Wu *et al.*, 2023	Zanini *et al.*, 2018
Cell culture / cell line	HBEC	A549-ACE2	nHTBE	ALOs	nHTBE	A549	A549	A549	Placenta hTSC	huh7
Virus	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2	IAV	IAV	IAV	IAV	ZIKV	ZIKV
# Cells	56798	42053	3077	12530	5045	15268	22908	49588	37062	1760

Dendrogram of hierarchical clustering



Cluster	Direction	Term name
Cluster a	down	Cell Cycle: • Meiotic cell cycle • DNA biosynthetic process • Mitotic nuclear division
Cluster b	up	Inflammatory Response: • Adaptive immune response • Leukocyte cell-cell adhesion • Interleukine-10 production
	down	Oxidative Phosphorylation: • Oxidative phosphorylation • Cellular respiration • Electron transport chain
Cluster c	up	Transport: • Nuclear migration • Vesicle targeting • Import into nucleus
	down	Transport: • Inorganic anion transport • Calcium ion import • Regulated exocytosis
Cluster d	up	Oxidative Phosphorylation: • Oxidative phosphorylation • Cellular respiration • Electron transport chain
	down	Inflammatory Response: • Adaptive immune response • Positive regulation of defense response • Regulation of T cell activation



CONCLUSION

Our results demonstrate that semantic categorization of gene sets enables a systematic comparison of host transcriptional responses across different viral infections. Despite differences between viruses, we identified conserved host cell reprogramming patterns, particularly the **upregulation of mitochondrial and energy metabolism pathways** and the suppression of differentiation-associated processes. Furthermore, virus-specific responses, such as alterations in mineral ion transport observed in Zika and Influenza infections or fatty acid metabolic changes, highlight distinct mechanisms of host manipulation. These findings illustrate that embedding-based semantic gene set analysis provides a powerful framework for integrating and comparing large transcriptomic datasets, facilitating the identification of conserved host responses and potential host-directed therapeutic targets.

References

Wu *et al.*, 2023, Nat Commun, 8;14(1)
Zanini *et al.*, 2018, Elife 16;7.
Ravindra *et al.*, 2021, PloS Biol., 19(3)
Nilsson-Payant *et al.*, 2021, J Virol 9;95(23)
Fiege *et al.*, 2020, bioRxiv
Karim *et al.*, 2025, Nat Commun, 10;16(1)
Roach *et al.*, 2024, Proc Natl Acad Sci U S A 30;121(31)
Russel *et al.*, 2018, Elife, 16;7
Wang *et al.*, 2020, mBio 14;11(1)
Rivera-Cardona, 2026, Proc Natl Acad Sci U S A 123(31)
Reffsin et al. 2025, Cell 189(1)
Keogh *et al.*, 2025, Comput Struct Biotechnol J

Funded by the European Union
www.project-appeal.eu
Funded by the European Union under GA No 101137311 and by the BMBF (project "SARS-CoV-2Dx", FKZ: 13N15745).
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.
SPONSORED BY THE Federal Ministry of Education and Research