

ELISA: A GENERATIVE AI AGENT FOR EXPRESSION-GROUNDED DISCOVERY IN SINGLE-CELL GENOMICS

Omar Coser

No Affiliation

Seeking Postdoctoral Position in Bioinformatics and AI

omarcoser10@gmail.com

ABSTRACT

We present **ELISA**, a retrieval-augmented AI agent for interpretable, hypothesis-driven exploration of single-cell RNA sequencing (scRNA-seq) data. ELISA enables natural-language querying of cell populations through a *query-conditioned retrieval framework* that explicitly integrates semantic biological priors with expression-derived evidence. In *semantic mode*, cluster level biological summaries are embedded using BioBERT to align user queries with ontology-supported annotations. *Hybrid mode* extends this approach by constructing a query-adaptive expression representation from semantically relevant clusters and combining it with scGPT-derived transcriptional embeddings, prioritizing cell populations that are both semantically relevant and transcriptionally coherent with the query intent. *scGPT mode* relies exclusively on transcriptional structure captured in the scGPT latent embedding space, emphasizing genes that dominantly shape expression-derived representations of retrieved clusters, independent of semantic annotations or curated biological knowledge. Finally, *discovery mode* contrasts dataset-specific expression signals with prior biological knowledge to surface context-shifted gene programs and generate cautious, data grounded hypotheses. By explicitly separating retrieval, expression evidence, and language-model interpretation, ELISA prioritizes transparency and reproducibility over speculative inference. The system is designed as a human-in-the-loop analytical tool that supports expert reasoning and hypothesis generation rather than fully autonomous biological discovery.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) enables genome-wide transcriptional profiling at single-cell resolution, as first demonstrated by early mRNA-seq approaches Tang et al. (2009). Since then, scRNA-seq has become a core tool in modern biology, supporting the systematic characterization of cell types, states, and developmental trajectories across tissues and conditions. This progress has been enabled by increasingly standardized analytical pipelines for preprocessing, clustering, and differential expression analysis Luecken & Theis (2019). Despite these advances, interpreting large single-cell atlases remains challenging due to biological heterogeneity, context-dependent cell states, and the difficulty of integrating statistical signals with rich prior biological knowledge.

In parallel, retrieval-augmented generation (RAG) has emerged as an effective framework for grounding large language models (LLMs) in external knowledge sources Lewis et al. (2020). LLMs have been shown to encode substantial biomedical knowledge Singhal et al. (2023), motivating their use in biological reasoning, annotation, and hypothesis exploration. Recent systems extend this paradigm to domain-specific AI agents for gene-set analysis and biomedical discovery Wang et al. (2025); Gao et al. (2024), as well as general-purpose biomedical agents Huang et al. (2025) and multi-agent scientific discovery frameworks Swanson et al. (2025). However, these approaches primarily operate over text and curated databases, limiting their ability to reason directly over high-dimensional experimental data such as scRNA-seq.

Concurrently, foundation models for single-cell biology such as scGPT learn expressive latent representations directly from transcriptomic data Cui et al. (2024). These models capture complex gene-gene dependencies and cellular structure, enabling powerful downstream analyses including cell embedding, label transfer, and perturbation modeling. However, expression-derived representations produced by such models are not designed to support semantic querying or interaction with free-form biological concepts expressed in natural language. As a result, despite major advances in both language-based biomedical reasoning and expression-based single-cell modeling, these two paradigms remain largely disconnected.

Our agent, ELISA (Figure 1), bridges this gap by integrating expression-derived embeddings with semantic retrieval and LLM-based interpretation for interactive single-cell discovery. Rather than retraining or modifying expression foundation models, ELISA treats scGPT as an expression-side representation that is explicitly combined with semantic embeddings derived from biological annotations and user queries. By separating data-driven evidence from prior biological knowledge and fusing them only at retrieval time, ELISA enables transparent exploration and cautious hypothesis generation grounded in both transcriptional structure and established biology. To our knowledge, scGPT embeddings have not previously been integrated with semantic language representations in a query-conditioned retrieval framework for single-cell genomics.

In particular, we make the following contributions:

- We propose a **cluster-centric representation** of single-cell data that integrates differential expression, pathway enrichment, and metadata into biologically interpretable textual summaries.
- We introduce a **hybrid retrieval framework** combining in a multimodal way semantic similarity in language embedding space with expression-derived structure captured by scGPT embeddings.
- We define a **discovery mode** that merges semantic and hybrid evidence to support hypothesis generation grounded in dataset-specific biology.

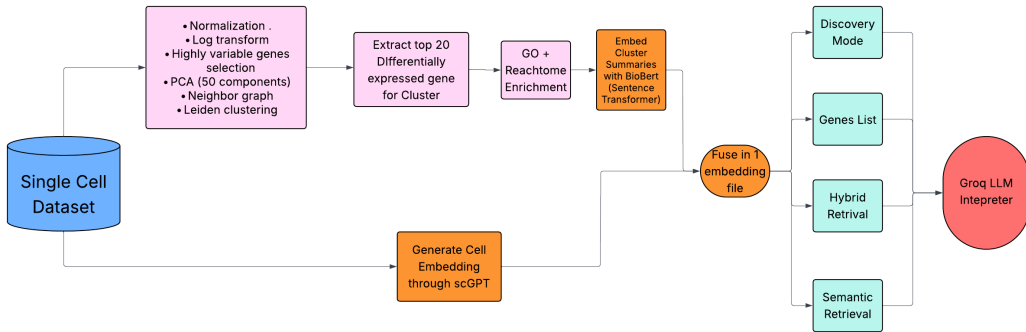


Figure 1: Overview of ELISA. Single-cell RNA-seq data are summarized at the cluster level and represented using semantic embeddings derived from biological annotations and expression-derived embeddings obtained from scGPT. Semantic, hybrid and scGPT retrieval support natural language querying, while discovery mode contrasts dataset evidence with prior biological knowledge to generate cautious, interpretable hypotheses, (Groq LLM use the llama-3.1-8b-instant model).

2 MATERIALS

Dataset. We evaluate ELISA on two publicly available scRNA-seq datasets. The first is a human airway atlas in which cells are annotated by patient condition (e.g., disease vs. control), laboratory of origin, and hierarchical cell-type labels Berg et al. (2025). The second dataset is a large-scale single-cell RNA-seq atlas of human immune cells spanning multiple tissues and donors. From this resource, we extracted a subset of immune populations, including macrophages, monocytes, mast cells, and microglia Vinsland et al. (2025).

Computational Resources All experiments have been performed on an A100 GPU with 80GB of Ram.

3 METHODS

Pre-processing and clustering. Starting from a gene-by-cell count matrix, we apply a standard scRNA-seq pipeline Luecken & Theis (2019): (i) library-size normalization and log-transformation; (ii) selection of highly variable genes (2,000–3,000); (iii) dimensionality reduction via PCA (30–50 PCs); (iv) kNN graph construction ($k = 15$ –30) and community detection (Leiden/Louvain; resolution 0.4–1.2); (v) cluster-specific marker detection using Wilcoxon rank-sum tests, yielding log fold-change ($\log FC$) and detection fractions inside vs. outside the cluster (pct_{in} , pct_{out}). All downstream representations are computed at the *cluster level*.

Cluster-centric evidence pack. For each cluster c_k , ELISA constructs a structured evidence object consisting of: (i) metadata summaries (cell counts, condition and batch composition, coarse/fine cell types); (ii) ranked marker genes with ($\log FC$, pct_{in} , pct_{out}); (iii) optional GO and Reactome enrichment terms. This evidence pack is the *only* information exposed to the LLM.

3.1 REPRESENTATION AND EMBEDDING CONSTRUCTION

ELISA builds two complementary cluster-level representations: semantic embeddings derived from biological annotations using BioBERT Lee et al. (2020), and expression-derived embeddings computed using scGPT Cui et al. (2024).

Semantic embeddings. Each cluster is summarized as a text description t_k capturing dominant cell types, pathway enrichments, and metadata. Semantic embeddings are computed as

$$\mathbf{s}_k = \text{BioBERT}(t_k) \in \mathbb{R}^{768}. \quad (1)$$

A query q is embedded as \mathbf{s}_q , and semantic similarity is measured by cosine similarity:

$$\text{sim}_{\text{sem}}(q, c_k) = \cos(\mathbf{s}_q, \mathbf{s}_k). \quad (2)$$

Expression-derived embeddings. Single cells are embedded using a pretrained scGPT encoder,

$$\mathbf{g}_i = \text{scGPT}(\mathbf{x}_i), \quad (3)$$

and aggregated at the cluster level by averaging:

$$\mathbf{g}_k = \frac{1}{|c_k|} \sum_{i \in c_k} \mathbf{g}_i. \quad (4)$$

Hybrid retrieval. Given a query q , we first retrieve the top- k clusters using semantic similarity. Marker genes from these clusters define a query-relevant gene set, which is embedded using scGPT to obtain a query expression vector:

$$\mathbf{g}_q = \text{mean}(\text{scGPT}(\text{top-}n \text{ genes})). \quad (5)$$

After ℓ_2 normalization, expression similarity is computed as

$$\text{sim}_{\text{expr}}(q, c_k) = \cos(\hat{\mathbf{g}}_q, \hat{\mathbf{g}}_k). \quad (6)$$

The final hybrid score is a convex combination:

$$\text{sim}_{\text{hyb}}(q, c_k) = \lambda \cdot \text{sim}_{\text{sem}}(q, c_k) + (1 - \lambda) \cdot \text{sim}_{\text{expr}}(q, c_k), \quad (7)$$

with $\lambda = 0.5$ unless otherwise stated. Gene-level information is introduced only *after* semantic retrieval, preserving a clear separation between language-based priors and expression-derived evidence.

3.1.1 GENE EVIDENCE PROCESSING

Gene-level evidence is entirely precomputed via differential expression analysis. Cell embeddings (scGPT) are used exclusively for cluster-level retrieval, while gene evidence is restricted to fixed cluster-specific marker sets. After retrieval, marker genes from the top- K clusters are pooled, filtered by expression salience (e.g., $\log FC$, pct_{in}), and ranked by maximum absolute $\log FC$ across retrieved clusters. Retrieval scores are not used for gene re-ranking.

Coverage is defined as the fraction of genes from curated canonical gene sets recovered among the retrieved markers. Top- M truncation is used as a robustness check.

3.1.2 INTERPRETATION ACROSS RETRIEVAL MODES

Semantic, scGPT-only, and hybrid retrieval operate over the same precomputed marker sets. Differences between modes arise solely from cluster selection: semantic retrieval emphasizes annotation-aligned signals, scGPT-only retrieval emphasizes expression-dominant structure, and hybrid retrieval integrates both without introducing new genes or altering gene-level scoring.

3.1.3 DISCOVERY MODE

In discovery mode, ELISA combines semantic and hybrid retrieval to extract clusters, marker genes, ontology terms, and metadata, which are passed to an LLM for staged interpretation: (i) summarization of dataset-derived signals, (ii) recall of known biology, (iii) comparison with expectations, and (iv) hypothesis generation. The LLM performs no retrieval or embedding, ensuring interpretability and reproducibility.

4 RESULTS AND DISCUSSION OF SEMANTIC VS HYBRID VS SCGPT RETRIEVAL

Across both epithelial- and immune-focused datasets, hybrid retrieval ($\lambda = 0.5$) consistently outperforms semantic-only ($\lambda = 1$) and expression-only scGPT retrieval ($\lambda = 0$), as measured by coverage of curated canonical gene sets (Tables 3 and ??). These results indicate that the three retrieval modes capture complementary biological signals, with hybrid retrieval providing the most complete and robust recovery of canonical transcriptional programs.

Semantic retrieval emphasizes ontology-aligned biology. Semantic retrieval prioritizes well-established, annotation-supported genes associated with conserved cellular identity and broadly characterized biological processes (Tables 3 and ??). This mode reliably recovers canonical stress-response chaperones (e.g., *HSPA1A/B*, *HSP90AA1*), lineage markers (e.g., *CIQA/C*, *PLXDC2*), and immune identity genes, yielding strong performance for queries closely aligned with curated pathway annotations. However, semantic retrieval exhibits systematic limitations for context-dependent and transient programs, including epithelial interferon signaling, antigen presentation, and cytokine responses, where coverage is often low or absent.

scGPT-only retrieval captures expression structure but lacks functional grounding. Expression-only retrieval using scGPT embeddings identifies clusters with coherent transcriptional states, performing well for innate-like signaling, inflammatory heterogeneity, and regulatory immune programs in Dataset 2 (Table ??). These results highlight the ability of scGPT to capture expression-driven similarities reused across immune lineages. Nevertheless, scGPT-only retrieval underperforms for queries requiring explicit functional interpretation, such as interferon responses, immune activation transitions, and cytokine signaling. The absence of semantic priors limits recovery of canonical stress and regulatory genes (e.g., *HSPA1A/B*, *DUSP1*, *ATF3*), resulting in lower and less stable coverage.

Hybrid retrieval achieves the most complete coverage. Hybrid retrieval consistently achieves the highest or near-highest coverage across both datasets (Tables 3 and ??, in Appendix). By integrating semantic similarity with expression-derived structure, hybrid retrieval preserves alignment with ontology-supported biology while expanding coverage to include regulatory and activation-associated programs that are weakly represented in semantic annotations alone. This advantage

is particularly evident for epithelial oxidative stress, cytokine signaling, antigen presentation, immune exhaustion, and interferon-stimulated programs, where hybrid retrieval recovers additional transcriptional regulators and signaling adaptors (e.g., *HSPB1*, *TREM2*, *ALOX5AP*, *KLF6*, *DUSP1*, *ATF3*). Importantly, hybrid retrieval does not degrade performance for well-annotated immune programs, matching or exceeding semantic coverage.

Implications for multimodal retrieval. Together, these results demonstrate that semantic and expression-based retrieval methods are complementary rather than redundant. Semantic retrieval anchors results in stable, ontology-supported knowledge, while scGPT captures latent transcriptional structure. Hybrid retrieval effectively unifies these perspectives, yielding the most consistent and biologically meaningful recovery of canonical gene programs across diverse epithelial and immune contexts.

Quantitative summary. A global summary of retrieval efficiency across all queries and both datasets is shown in Figure 2. Consistent with Tables 3 and ??, hybrid retrieval achieves the highest mean coverage with reduced variance across queries, indicating improved robustness and more reliable recovery of context-dependent functional programs.

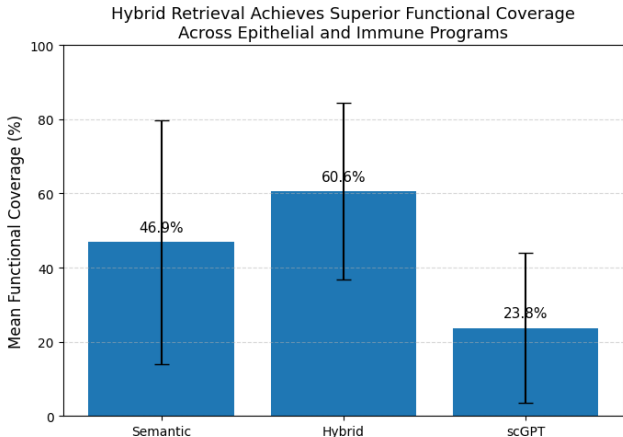


Figure 2: Mean functional coverage across all queries in Dataset 1 and Dataset 2 for semantic-only ($\lambda = 1$), hybrid ($\lambda = 0.5$), and scGPT-only ($\lambda = 0$) retrieval. Error bars denote standard deviation across queries. Hybrid retrieval achieves the highest average coverage with reduced variance across queries.

Table 1: Dataset 1, Semantic vs Hybrid.

Query	Mode	Top Clusters	Key Gene Evidence	Cov
1. Epithelial Stress Responses				
Epithelial sponse	stress re-Semantic	3, 2, 6, 8, 13	HSPA1B, HSPA1A, HSP90AA1, HSPB1, RPS4Y1, HBA1, HBA2, HBZ, C1QC, C1QB, CD74, PLXDC2	30
Epithelial sponse	stress re-Hybrid	3, 2, 6, 8, 13	HSPA1B, HSPA1A, HSP90AA1, SELENOP, HBA1, HBA2, HBZ, RPS26, C1QC, DAB2, TTR, CD74, PLXDC2	30
Epithelial sponse	stress re-scGPT	6, 1, 17, 3	SLPI, SCGB3A1, KRT19, KRT17, MIR205HG, JUNB, HLA-DRA, CD74, CCL5	0

Query	Mode	Top Clusters	Key Gene Evidence	
Oxidative stress signaling in epithelial cells	Semantic	3, 15, 2, 11, 16	HSPA1B, HSPA1A, SLC9A9, RPS4Y1, PLXDC2, S100A11, FRMD4A, SLC2A1, CNST, LTBP1, LIMS1, GPNMB, NEAT1	40
Oxidative stress signaling in epithelial cells	Hybrid	3, 15, 2, 11, 16	HSPA1B, HSPA1A, HSPB1, TREM2, ALOX5AP, CNST, LIMS1, GPNMB, NEAT1	50
Oxidative stress signaling in epithelial cells	scGPT	6, 1, 3, 15, 17	SLPI, SCGB3A1, KRT19, KRT17, FOS, HSPB1, CCL5, PTPRC, SRGN, CD74	10
Immune-like Signaling in Epithelial Cells				
Immune-like signaling in epithelial cells	Semantic	15, 1, 11, 2, 16	DAB2, C1QA, FCGRT, PLXDC2, S100A11, TREM2, RPS4Y1, HSPA1A, HSPA1B, SLC2A1, GPNMB, NEAT1	80
Immune-like signaling in epithelial cells	Hybrid	15, 1, 11, 2, 16	CNST, LTBP1, LIMS1, LGMN, C1QA, FCGRT, PLXDC2, S100A11, TREM2, HSPA1A, HSPA1B, GPNMB, NEAT1, KLF6	80
Immune-like signaling in epithelial cells	scGPT	6, 1, 15, 17, 3	SLPI, SCGB3A1, KRT19, KRT17, CD9, PTPRC, CCL5, HLA-DRA, CD74	10
Interferon signaling in airway epithelial cells	Semantic	12, 15, 14, 13, 8	Nan	0
Interferon signaling in airway epithelial cells	Hybrid	12, 15, 14, 8, 13	TUBB, TUBA1B, H2AZ1, STMN1, RPS26, RPL39, TTR, CD74, PLXDC2	30
Interferon signaling in airway epithelial cells	scGPT	6, 15, 1, 3, 17	CCL5, PTPRC, SRGN	0
Antigen Presentation in Epithelial Cells				
Antigen presentation in epithelial	Semantic	18, 1, 15, 11, 20	Nan	0
Antigen presentation in epithelial	Hybrid	18, 1, 15, 11, 20	LGMN, C1QA, FCGRT, PLXDC2, S100A11, TREM2	50
Antigen presentation in epithelial cells	scGPT	1, 6, 15, 17, 8	SLPI, SCGB3A1, PTPRC, CCL5, HLA-DRA, HLA-DPB1, FYN	10
MHC-II expression in epithelial cells	Semantic	2, 14, 18, 1, 11	Nan	0
MHC-II expression in epithelial cells	Hybrid	2, 14, 18, 1, 11	H2AZ1, TUBA1B, S100A6, TIMP3, C1QA, CD14, PLXDC2, S100A11	30
MHC-II expression in epithelial cells	scGPT	6, 3, 17	HLA-DRA, HLA-DPB1, CD74, HLA-DQA1	20
Inter-epithelial Immune Signaling				
Inter-epithelial immune signaling	Semantic	11, 15, 13, 18, 8	PLXDC2, S100A11, TREM2, CNST, LIMS1, RGS18, CD74, ARHGAP15, GATA2, S100A6, RPS26, C1QC, DAB2	80
Inter-epithelial immune signaling	Hybrid	11, 15, 18, 8, 13	PLXDC2, SLC8A1, S100A11, CNST, LTBP1, LIMS1, ARHGAP15, GATA2, S100A6, RPS26, C1QC, DAB2, TTR, CD74	80

Query	Mode	Top Clusters	Key Gene Evidence
Inter-epithelial immune signaling	scGPT	6, 1, 3, 15, 17	SLPI, SCGB3A1, KRT19, CD9, HSPB1, CCL5, PTPRC, ARHGAP15, CD74
Cytokine signaling within epithelial	Semantic	15, 20, 16, 13, 11	CD52, TRBC1, IL7R, GPNMB, NEAT1, KLF6, TTR, CD74, PLXDC2, S100A11, DOCK4
Cytokine signaling within epithelial	Hybrid	15, 20, 16, 11, 1	IL7R, NFKBIA, KLF6, FOS, DOCK4, TREM2, LIMS1, RGS18, LGMN, C1QA
Cytokine signaling within epithelial cells	scGPT	6, 1, 3, 15, 17	SLPI, SCGB3A1, KRT19, KRT17, CD9, FOS, CCL5, PTPRC, HLA-DRA

5 DISCOVERY MODE REVEALS CONTEXT-SHIFTED BIOLOGICAL PROGRAMS BEYOND ANNOTATION-DRIVEN RETRIEVAL.

Tables 2 and 4 illustrate the qualitative behavior of ELISA’s discovery mode across epithelial and immune-centric queries. In contrast to semantic or hybrid retrieval, which prioritize ontology alignment or expression similarity to predefined concepts, discovery mode explicitly surfaces clusters where strong gene-level evidence appears in unexpected cellular or functional contexts.

In epithelial-focused queries (Table 2), discovery mode highlights antigen presentation, hypoxic signaling, and immune-associated transcriptional programs in epithelial populations not canonically annotated as immune. The presence of immune markers (e.g., *HLA-DRA*, *CD74*, *TYROBP*) alongside ciliary or secretory genes suggests potential stress-induced or context-dependent immune–epithelial coupling, rather than annotation errors. Importantly, ELISA does not assert novel functions, but instead formulates cautious hypotheses grounded in observed expression patterns.

In immune-focused queries (Table 4, in appendix), discovery mode consistently identifies transcriptional modules associated with development, cytoskeletal remodeling, and generic stress responses across mature immune populations. The recurrence of developmental transcription factors (e.g., *GATA2*, *MEIS2*, *SOX4*), cytoskeletal regulators (e.g., *TAGLN2*, *ITGA4*), and stress-response genes (e.g., *HSPA1A/B*, *ATF3*) suggests that immune activation may involve partial reuse of broadly conserved regulatory programs rather than strictly immune-specific pathways.

Together, these tables demonstrate that discovery mode complements semantic and hybrid retrieval by emphasizing hypothesis generation over retrieval accuracy. By explicitly contrasting dataset evidence with established biology, ELISA supports exploratory analysis while maintaining interpretability and epistemic caution—an essential requirement for trustworthy biological discovery.

Table 2: Discovery mode Dataset 1

Query	Mode	Top Clusters	Key Gene Evidence	Unexpected / Context-shifted	Formulated Hypotheses
1. Antigen Presentation in Unexpected Contexts					
Unexpected antigen presentation in non-immune cells	Discovery	10, 8, 16, 0, 11	CD3D, CCL5, PTPRC, HLA-DRA, TYROBP	DNAH12, CFAP299, LRR1Q1 in epithelial cluster 0	Ciliary-associated genes may participate in immune-related signaling or reflect epithelial stress–immune coupling
2. Hypoxic and Stress Responses					
Hypoxic signaling in airway basal cells	Discovery	12, 5, 2, 4, 14	FAM155A, ARHGAP15, KRT19	IL32, HSPB1, LRR1Q1, Z93241.1, GASS	Basal epithelial and immune-associated genes may jointly contribute to hypoxia-adaptive programs
3. Immune–Epithelial Crosstalk					
Crosstalk between immune and secretory epithelial cells	Discovery	2, 6, 7, 14, 11	KRT19, SLPI, CD74, MIR205HG	CD9, Z93241.1, GASS, AQP3	Secretory epithelial cells may actively participate in immune modulation beyond canonical roles

6 CONCLUSION

ELISA is a retrieval-augmented agent designed to support exploratory analysis of large-scale single-cell transcriptomic data by integrating semantic understanding with expression-derived biological

evidence. In contrast to traditional annotation-centric tools, ELISA enables free-form biological querying while grounding results in transcriptionally coherent cell populations and interpretable gene-level signals.

By combining domain-adapted language embeddings with scGPT-based expression representations, ELISA supports multiple retrieval modes that address distinct analytical needs. Semantic retrieval provides efficient alignment with ontology-supported annotations, expression-based retrieval captures latent transcriptional structure, and hybrid retrieval integrates both to improve sensitivity to context-dependent functional programs. This multimodal design allows ELISA to recover biologically meaningful states that may be weakly annotated, transient, or shared across cell types.

Across diverse epithelial and immune queries, hybrid retrieval consistently achieves broader and more robust coverage of curated canonical gene programs, demonstrating the value of integrating prior biological knowledge with data-driven expression structure. Importantly, ELISA exposes retrieval results alongside gene-level evidence, enabling users to critically assess biological relevance rather than relying on opaque annotations alone.

More broadly, ELISA illustrates how agent-based architectures can bridge natural language interfaces and high-dimensional biological data in a transparent and scalable manner. By anchoring free-text queries to interpretable transcriptional evidence, ELISA provides a practical framework for hypothesis generation, cross-dataset exploration, and mechanistic reasoning in single-cell genomics. Importantly, ELISA is not intended as a fully autonomous discovery system, but as a human-in-the-loop analytical tool that supports expert reasoning and cautious interpretation rather than speculative inference. Qualitative review by a quantitative biologist confirmed that discovery-mode outputs consistently highlighted biologically plausible, context-dependent gene programs while maintaining appropriate epistemic caution. We view this approach as a step toward interactive, evidence-aware systems biology tools that operate at the interface of language, data, and biological insight.

7 AI-ASSISTANCE

ChatGPT was used solely to improve the clarity and quality of the text and to correct typographical and grammatical errors

REFERENCES

- Marijn Berg, Lisette Krabbendam, Esmee K van der Ploeg, Menno van Nimwegen, Tjeerd van der Veer, Martin Banchero, Orestes A Carpaij, Remco Hoogenboezem, Maarten van den Berge, Eric Bindels, et al. Evidence for altered immune-structural cell crosstalk in cystic fibrosis revealed by single cell transcriptomics. *Journal of Cystic Fibrosis*, 2025.
- Christine M Bojanowski, Shiping Lu, and Jay K Kolls. Mucosal immunity in cystic fibrosis. *The Journal of Immunology*, 207(12):2901–2912, 2021.
- Junyue Cao, Diana R O’Day, Hannah A Pliner, Paul D Kingsley, Mei Deng, Riza M Daza, Michael A Zager, KA Aldinger, R Blecher-Gonen, F Zhang, et al. A human cell atlas of fetal gene expression. *science* 370, eaba7721, 2020.
- Gianni Carraro, Justin Langerman, Shan Sabri, Zareeb Lorenzana, Arunima Purkayastha, Guangzhu Zhang, Bindu Konda, Cody J Aros, Ben A Calvert, Aleks Szymaniak, et al. Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. *Nature medicine*, 27(5):806–814, 2021.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.
- Juliette Eugène, Nicolas Jouand, Kathleen Ducoin, Delphine Dansette, Romain Oger, Cécile Deleine, Edouard Leveque, Guillaume Meurette, Juliette Podevin, Tamara Matysiak, Jaafar Ben-nouna, Stéphane Bezieau, Christelle Volteau, Wassila El Alami Thomas, Jérôme Chetritt, Olivier Kerdraon, Pierre Fourquier, Emilie Thibaudeau, Frédéric Dumont, Jean-François Mosnier, Claire Toquet, Anne Jarry, Nadine Gervois, and Céline Bossard. The inhibitory receptor cd94/nkg2a

- on cd8+ tumor-infiltrating lymphocytes in colorectal cancer: a promising new druggable immune checkpoint in the context of hlae/2m overexpression. *Modern Pathology*, 33(3):468–482, 2020. ISSN 0893-3952. doi: <https://doi.org/10.1038/s41379-019-0322-9>. URL <https://www.sciencedirect.com/science/article/pii/S0893395222009164>.
- Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.
- Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jae-woo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.
- Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.
- Manuella Lévêque, Sophie Le Trionnaire, Paola Del Porto, and Corinne Martin-Chouly. The impact of impaired macrophage functions in cystic fibrosis disease progression. *Journal of Cystic Fibrosis*, 16(4):443–453, 2017. ISSN 1569-1993. doi: <https://doi.org/10.1016/j.jcf.2016.10.011>. URL <https://www.sciencedirect.com/science/article/pii/S1569199316306464>.
- Sorcha O’Byrne, Natalina Elliott, Siobhan Rice, Gemma Buck, Nicholas Fordham, Catherine Garnett, Laura Godfrey, Nicholas T Crump, Gary Wright, Sarah Inglott, et al. Discovery of a cd10-negative b-progenitor in human fetal life identifies unique ontogeny-related developmental programs. *Blood, The Journal of the American Society of Hematology*, 134(13):1059–1071, 2019.
- Jonas C Schupp, Sara Khanal, Jose L Gomez, Maor Sauler, Taylor S Adams, Geoffrey L Chupp, Xiting Yan, Sergio Poli, Yujiao Zhao, Ruth R Montgomery, et al. Single-cell transcriptional archetypes of airway inflammation in cystic fibrosis. *American journal of respiratory and critical care medicine*, 202(10):1419–1429, 2020.
- Michal Shteinberg, Iram J Haq, Deepika Polineni, and Jane C Davies. Cystic fibrosis. *The Lancet*, 397(10290):2195–2211, 2021. ISSN 0140-6736. doi: [https://doi.org/10.1016/S0140-6736\(20\)32542-3](https://doi.org/10.1016/S0140-6736(20)32542-3). URL <https://www.sciencedirect.com/science/article/pii/S0140673620325423>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.
- Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.
- Yasuo Tohma, Tetsumori Yamashima, and Junkoh Yamashita. Immunohistochemical localization of cell adhesion molecule epithelial cadherin in human arachnoid villi and meningiomas. *Cancer research*, 52(7):1981–1987, 1992.

Nadine van Montfoort, Linda Borst, Michael J Korrer, Marjolein Sluijter, Koen A Marijt, Saskia J Santegoets, Vanessa J van Ham, Ilina Ehsan, Pornpimol Charoentong, Pascale André, et al. Nkg2a blockade potentiates cd8 t cell immunity induced by cancer vaccines. *Cell*, 175(7):1744–1755, 2018.

Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573, 2017.

Elin Vinsland, Sergio Marco Salas, Ivana Kapustová, Lijuan Hu, Simone Webb, Xiaofei Li, Xiaoling He, Mats Nilsson, Muzlifah Haniffa, Roger Barker, et al. Cell atlas of the developing human meninges reveals a dura origin of meningioma. *bioRxiv*, 2025.

Anthony Z Wang, Jay A Bowman-Kirigin, Rupen Desai, Liang-I Kang, Pujan R Patel, Bhuvic Patel, Saad M Khan, Diane Bender, M Caleb Marlin, Jingxian Liu, et al. Single-cell profiling of human dura and meningioma reveals cellular meningeal landscape and insights into meningioma immune response. *Genome medicine*, 14(1):49, 2022.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025.

A APPENDIX

B TABLE DATASET TWO

Table 3: Dataset 2, Semantic vs Hybrid.

Query	Mode	Top Clusters	Key Gene Evidence	Cov
Immune Activation State Heterogeneity Across Immune Cell Populations				
Immune activation state heterogeneity across immune cell populations	Semantic	18, 3, 16, 1, 11	ARHGAP15, GATA2, TIMP3, HSPA1B, HSPA1A, SLC9A9, GPNMB, NEAT1, KLF6, LGMN, C1QA, DAB2, PLXDC2, S100A11, DOCK4	60
Immune activation state heterogeneity across immune cell populations	Hybrid	18, 3, 16, 1, 11	ARHGAP15, GATA2, S100A6, GIHCG, SLC24A3, MEIS2, ITGA4, TAGLN2, SOX4, HSPA1B, HSPA1A, HSP90AA1, RPS4Y1, DNAJB1, STAB1, HSPB1, JUN, GPNMB, NEAT1, SLC2A1, KLF6, FOS, DUSP1, ATF3, SH3BP5, ARL4C, SAMD4A, LGMN, C1QA, DAB2, FCGRT, BLVRB, PLTP, VSIG4, A2M, C1QB, CD14, PLXDC2, SLC8A1, S100A11, FRMD4A, DOCK4, TREM2, ALOX5AP, CAPG, LIMS1, C3	100
Immune activation state heterogeneity across immune cell populations	scGPT	16, 17, 19, 8, 11	GPNMB, NEAT1, RUNX1, LYVE1, C1QC, DAB2, PLXDC2, S100A11	10
Transcriptional programs associated with activated versus resting immune cells	Semantic	3, 16, 18, 19, 1	HSPA1B, HSPA1A, GPNMB, NEAT1, ARHGAP15, ITGA4, RUNX1, LYVE1, LGMN, C1QA	50
Transcriptional programs associated with activated versus resting immune cells	Hybrid	3, 16, 18, 19, 1	HSPA1B, HSPA1A, HSP90AA1, GPNMB, NEAT1, KLF6, ARHGAP15, GATA2, S100A6, FOXO3, LGMN, C1QA, DAB2	60
Transcriptional programs associated with activated versus resting immune cells	scGPT	16, 17, 19	GPNMB, NEAT1, RUNX1, LYVE1, FOXO3	10
Shared Transcriptional Programs Between Innate and Adaptive Immune Cells				
Shared transcriptional programs between innate and adaptive immune cells	Semantic	3, 18, 19, 16, 8	HSPA1B, HSPA1A, HSP90AA1, ARHGAP15, ITGA4, RUNX1, LYVE1, GPNMB, SLC2A1, RPS26, C1QC	90
Shared transcriptional programs between innate and adaptive immune cells	Hybrid	3, 8, 16, 18, 19	HSPA1B, HSPA1A, SLC9A9, ARHGAP15, GATA2, RUNX1, LYVE1	60

Shared transcriptional programs between innate and adaptive immune cells	scGPT	16, 17, 19, 18	GPNMB, SLC2A1, RUNX1, LYVE1, ARHGAP15, GATA2	60
Innate-like signaling pathways expressed in adaptive immune populations	Semantic	3, 18, 2	HSPA1B, HSPA1A, SLC9A9, MRC1, ARHGAP15, GATA2, TIMP3, RPS4Y1	40
Innate-like signaling pathways expressed in adaptive immune populations	Hybrid	3, 18, 2, 8, 15	HSPA1B, HSPA1A, HSP90AA1, HSPB1, ARHGAP15, ITGA4, GATA2, MEIS2, RPS26, C1QC, RPL39, CNST, LIMS1, RGS18, PBX1	60
Innate-like signaling pathways expressed in adaptive immune populations	scGPT	17, 19, 8, 18	RUNX1, LYVE1, C1QC, ARHGAP15, GATA2	50
Immune Dysfunction, Exhaustion, and Regulation				
Transcriptional signatures of immune exhaustion or dysfunction	Semantic	20, 1, 19, 7, 0	CD52, TRBC1, IL7R, LGMN, C1QA, DAB2, RUNX1, LYVE1, A2M, CD74, VIM, HLA-DRA, FOS, RNASE1, FOLR2	30
Transcriptional signatures of immune exhaustion or dysfunction	Hybrid	20, 1, 19, 0, 3	CD52, TRBC1, ID2, LGMN, C1QA, DAB2, RUNX1, LYVE1, RPS4Y1, FOS, RNASE1, FOLR2, HSPA1B, HSPA1A, SLC9A9	50
Transcriptional signatures of immune exhaustion or dysfunction	scGPT	16, 17, 11, 19, 20	GPNMB, RUNX1, LYVE1, KLF6, TREM2	30
Regulatory immune programs beyond canonical Treg annotations	Semantic	18, 11, 10, 15, 16	ARHGAP15, GATA2, S100A6, PLXDC2, SLC8A1, S100A11, SRGN, RNASE1, LYVE1, CNST, LTBP1, LIMS1, GPNMB, NEAT1, SLC2A1	80
Regulatory immune programs beyond canonical Treg annotations	Hybrid	18, 11, 15, 16, 19	ARHGAP15, GATA2, TIMP3, S100A6, PLXDC2, SLC8A1, S100A11, FRMD4A, CNST, LTBP1, LIMS1, RGS18, GPNMB, NEAT1, SLC2A1, KLF6, RUNX1, LYVE1, RPS4Y1, FOXO3	90
Regulatory immune programs beyond canonical Treg annotations	scGPT	16, 17, 18, 19	GPNMB, SLC2A1, ARHGAP15, GATA2	40
Interferon and Inflammation Signaling				
Interferon-stimulated gene programs across immune cell types	Semantic	18, 3, 16, 2, 13	ARHGAP15, GATA2, TIMP3, S100A6, HSPA1B, HSPA1A, SLC9A9, MRC1, GPNMB, NEAT1, SLC2A1, KLF6, RPS4Y1, SELENOP, TTR, CD74, PLXDC2, RPL41	80
Interferon-stimulated gene programs across immune cell types	Hybrid	18, 3, 16, 2, 13	ARHGAP15, GATA2, S100A6, HSPA1B, HSPA1A, HSP90AA1, GPNMB, NEAT1, KLF6, HSPB1, TTR, CD74, PLXDC2	90

Interferon-stimulated gene programs across immune cell types	scGPT	16, 17, 19, 18, 8	GPNMB, NEAT1, SLC2A1, KLF6, FOS, DUSP1, ATF3	40
Inflammatory signaling heterogeneity within immune populations	Semantic	3, 18, 16, 2, 8	HSPA1B, HSPA1A, SLC9A9, ARHGAP15, GATA2, TIMP3, GPNMB, NEAT1, KLF6, RPS4Y1, RPS26, C1QC, DAB2	80
Inflammatory signaling heterogeneity within immune populations	Hybrid	3, 18, 16, 2, 8	HSPA1B, HSPA1A, HSP90AA1, HSPB1, ARHGAP15, GATA2, GPNMB, NEAT1, RPS26, C1QC	80
Inflammatory signaling heterogeneity within immune populations	scGPT	16, 17, 19, 8, 18	GPNMB, NEAT1, KLF6, C1QC, DAB2, ARHGAP15	60

Table 4: Discovery mode Dataset 2

Query	Mode	Top Clusters	Key Gene Evidence	Unexpected / Context-shifted	Formulated Hypotheses	
1. Developmental Program Reactivation						
Reactivation of developmental transcriptional programs during immune responses	Discovery	3, 18, 16	GATA2, MEIS2, SOX4, JUN, KLF6, GPNMB	Developmental transcription factors and early-response regulators expressed in mature immune populations	Immune activation may transiently reuse developmental transcriptional circuits to support plasticity, stress adaptation, and functional state transitions rather than lineage reversion	
2. Structural and Cytoskeletal Regulation						
Structural and cytoskeletal gene programs associated with immune regulation	Discovery	3, 18, 11	TAGLN2, ARHGAP15, DOCK4, VIM	ITGA4, CAPG	Cytoskeletal remodeling and adhesion programs enriched beyond canonical motility or activation contexts	Structural and cytoskeletal regulators may actively participate in immune signaling, migration plasticity, and cell-cell interaction remodeling
3. Stress and Transcriptional Rewiring						
Stress-associated transcriptional rewiring during immune activation	Discovery	3, 16, 8	HSPA1A/B, ATF3, DUSP1, FOS	HSP90AA1	Stress-response and immediate-early genes broadly activated across immune clusters	Immune responses may leverage generic stress-response transcriptional modules as a regulatory layer for rapid functional adaptation

B.1 THRESHOLDS AND HYPERPARAMETERS

All parameters are fixed *a priori*. Markers are selected using Wilcoxon tests with $p_{\text{adj}} < 0.05$, retaining the top $K = 20$ genes per cluster by $|\log FC|$. GO and Reactome enrichments are retained at $p_{\text{adj}} < 0.05$ (up to 10 terms each). Metadata summaries include only categorical fields with 2–20 levels, reporting categories representing at least 10% of cluster cells.

C FUTURE PERSPECTIVE

Future work will focus on strengthening the biological validation of ELISA by more systematically integrating ontology- and program-based evaluation strategies. In addition to Gene Ontology-based analyses, we plan to incorporate curated transcription factor-target databases such as TRRUST, enabling quantitative assessment of regulatory program recovery through enrichment of TF target genes among retrieved evidence sets. This will allow evaluation of context-dependent transcriptional programs beyond manually curated gene lists. In particular, incorporating hierarchical GO consistency, TF-target enrichment, and enrichment stability across retrieval modes would provide a more principled validation of functional coherence. We also plan to extend the current λ -ablation analysis by exploring adaptive or query-specific weighting strategies, allowing the balance between semantic priors and expression-derived evidence to be learned or dynamically adjusted rather than fixed. Finally, an important direction is to evaluate ELISA across additional datasets and perturbation settings, including cross-condition and cross-species analyses, to assess robustness, generalizability, and sensitivity to annotation sparsity. Together, these extensions would further improve the interpretability, validation rigor, and applicability of ELISA as a human-in-the-loop discovery tool for single-cell genomics.

C.1 QUANTITATIVE BIOLOGIST VALIDATION

Dataset 1 Berg et al. (2025) The canonical gene sets listed above were curated from well-established literature on epithelial stress, immune-epithelial interactions, and inflammatory signaling in the airway, and are not derived from the retrieval outputs themselves. Stress-associated chaperones and regulators (e.g., HSPA1A/B, HSP90AA1, HSPB1, HMOX1, TXNIP) capture conserved cellular responses to proteotoxic, oxidative, and inflammatory stress, while immune and antigen-presentation markers (e.g., HLA-DRA/DRB1, CD74, FCGRT, CIITA, C1QA/C) reflect well-characterized epithelial programs associated with immune sensing and immune-epithelial crosstalk. Interferon-stimulated genes (e.g., IFITM1, ISG15, MX1, OAS1, STAT1, IRF7) and cytokine-response regulators (e.g., NFKBIA, FOS, JUN, TNFAIP3, SOCS3) represent canonical antiviral and inflammatory pathways known to be context-dependent in epithelial cells.

Coverage values therefore quantify how effectively retrieval recovers biologically established markers, rather than defining ground truth. The consistent gains of hybrid retrieval for interferon signaling, antigen presentation, and cytokine responses indicate that integrating expression-derived structure with semantic priors improves sensitivity to transcriptional programs that are weakly represented in ontology-driven annotations but are well supported by gene-level evidence in the data.

- Epithelial stress response: HSPA1B, HSPA1A, HSP90AA1, TREM2, ALOX5AP, CNST, LIMS1, GPNMB, NEAT1, DAB2
- Oxidative stress signaling in epithelial cells: HSPA1A, HSPA1B, HSPB1, HMOX1, SOD2, TXNIP, NFKBIA, SLC2A1, ALOX5AP, GPNMB
- Immune-like signaling in epithelial cells: C1QA, FCGRT, CD74, TREM2, PLXDC2, DAB2, HSPA1A, HSPA1B, GPNMB, KLF6
- Interferon signaling in airway epithelial cells: IFITM1, ISG15, MX1, OAS1, STAT1, IRF7, HLA-DRA, CD74, PLXDC2, TTR
- Antigen presentation in epithelial: HLA-DRA, HLA-DRB1, CD74, LGMN, FCGRT, CTSS, CIITA, PLXDC2, TREM2, S100A11
- MHC-II expression in epithelial cells: HLA-DRA, HLA-DRB1, CD74, CIITA, CTSS, LGMN, FCGRT, C1QA, PLXDC2, S100A11

- Inter-epithelial immune signaling: CD74,C1QA,C1QB,PLXDC2,S100A11,TREM2,DAB2,ARHGAP15,GATA2,LTBP1
- Cytokine signaling within epithelial, IL7R, NFKBIA,FOS,JUN,CXCL8,IL6,TNFAIP3,SOCS3,CCL2,CXCL2

Gene has been selected by biologist knowledge and by the following work: Berg et al. (2025); Carraro et al. (2021); Schupp et al. (2020); Shteinberg et al. (2021); Bojanowski et al. (2021); Eugène et al. (2020); van Montfoort et al. (2018); Lévêque et al. (2017)

Dataset 2 Vinsland et al. (2025) The canonical gene sets for Dataset 2 capture well-established transcriptional programs underlying immune activation, functional state transitions, and regulatory heterogeneity across innate and adaptive immune populations. Immediate-early and stress-response regulators (e.g., HSPA1A/B, JUN, FOS, ATF3, DUSP1) are known hallmarks of immune activation and rapid transcriptional remodeling, while lineage-associated and regulatory factors (e.g., RUNX1, GATA2, ARHGAP15, ITGA4, LYVE1) reflect conserved immune identity, trafficking, and differentiation programs. Markers such as C1QA/C, GPNMB, LGMN, and DAB2 are repeatedly implicated in myeloid activation, immune regulation, and tissue-associated immune states, whereas interferon- and inflammation-associated genes (e.g., HSP90AA1, HSPB1, NEAT1, KLF6, CD74) represent canonical antiviral and inflammatory responses shared across immune cell types.

As in Dataset 1, coverage values measure recovery of biologically established transcriptional programs rather than discovery of novel markers. The increased coverage achieved by hybrid retrieval for activation heterogeneity, immune dysfunction, regulatory programs beyond canonical Treg annotations, and interferon signaling indicates improved sensitivity to broadly reused and context-dependent immune programs that are not fully captured by ontology-aligned semantic descriptions alone. In contrast, cases where semantic retrieval performs comparably (e.g., shared innate–adaptive programs and inflammatory signaling) suggest alignment with well-annotated immune pathways, highlighting the complementary strengths of semantic and expression-informed retrieval.

- Immune activation state heterogeneity across immune cell population:HSPA1A, HSPA1B, JUN, FOS, ATF3, DUSP1,C1QA, GPNMB, LGMN, KLF6
- Transcriptional programs associated with activated versus resting immune cells:HSPA1A, HSPA1B, JUN, FOS, ATF3, DUSP1, C1QA, GPNMB, LGMN, KLF6
- Shared transcriptional programs between innate and adaptive immune cells: HSPA1A, HSPA1B,RUNX1, GATA2,ARHGAP15, ITGA4,C1QC, GPNMB,SLC2A1, LYVE1
- Innate-like signaling pathways expressed in adaptive immune populations: HSPA1A, HSPA1B,RUNX1, GATA2,ARHGAP15, ITGA4,C1QC, GPNMB,SLC2A1, LYVE1
- Transcriptional signatures of immune exhaustion or dysfunction,HSPA1A, HSPA1B,RUNX1, GATA2,ARHGAP15, ITGA4,C1QA,GPNMB,SLC2A1,LYVE1
- Regulatory immune programs beyond canonical Treg annotations,ARHGAP15, GATA2,RUNX1,LYVE1,GPNMB,SLC2A1,C1QA,ITGA4,KLF6,FOX3
- Interferon-stimulated gene programs across immune cell types, HSPA1A,HSPA1B,HSP90AA1,HSPB1,GPNMB,NEAT1,SLC2A1,KLF6,CD74,TTR
- Inflammatory signaling heterogeneity within immune populations, HSPA1A,HSPA1B,HSP90AA1,HSPB1,GPNMB,NEAT1,KLF6,C1QC,DAB2,ARHGAP15

Gene have been selected through biologist knowledge and the following paper: Vinsland et al. (2025); Wang et al. (2022); Villani et al. (2017); O’Byrne et al. (2019); Cao et al. (2020); Tohma et al. (1992)

C.2 PROMPT DATASET 1

Semantic: Epithelial stress response.

Based on the provided dataset evidence, I will analyze the clusters related to the epithelial stress response. Cluster 3 has a semantic similarity of 0.22780147194862366 and contains 667 cells. The metadata indicates that these cells are primarily Myeloid cells, specif-

ically Macrophages Monocyte- like (94.60269865067467(HSPA1B, HSPA1A, HSP90AA1, HSPB1), which are often upregulated in response to stress. However, it is essential to note that the logFC values are NaN, indicating that the fold change is not available for these genes. Cluster 2 has a semantic similarity of 0.1988079845905304 and contains 958 cells. The metadata indicates that these cells are also primarily Myeloid cells, specifically Macrophages- iron metabolism (85.07306889352818proteins (HSPA1A, HSPA1B, RPS4Y1), which are similar to those found in Cluster 3. Cluster 6 has a semantic similarity of 0.18664202094078064 and contains 370 cells. The metadata indicates that these cells are primarily Myeloid cells, specifically Monocytes (85.67567756756757shows that these cells express genes involved in hemoglobin production (HBA1, HBA2, HBZ), which is not directly related to the epithelial stress response. Cluster 8 has a semantic similarity of 0.1846211552619934 and contains 342 cells. The metadata indicates that these cells are primarily Myeloid cells, specifically Monocytes (82.16374269005848express genes involved in the complement system (C1QC, C1QB), which is not directly related to the epithelial stress response. Cluster 13 has a semantic similarity of 0.17829006910324097 and contains 165 cells. The metadata indicates that these cells are primarily Myeloid cells, specifically ChP Microglia (99.39393939393939genes involved in the immune response (CD74, PLXDC2), which is not directly related to the epithelial stress response. In conclusion, Cluster 3 and Cluster 2 show some relevance to the epithelial stress response due to the expression of heat shock proteins. However, the lack of logFC values for these genes and the presence of other genes not directly related to stress response in other clusters limit the conclusion. Further analysis is required to confirm the relevance of these clusters to the epithelial stress response.

Hybrid: Epithelial stress response.

Based on the provided dataset evidence, I will analyze the clusters related to the epithelial stress response. Cluster 3 has a high hybrid similarity score of 0.8453 and is characterized by the presence of genes involved in stress response, such as HSPA1B, HSPA1A, and HSP90AA1. These genes are typically upregulated in response to cellular stress, suggesting that this cluster may be involved in the epithelial stress response. The cluster also contains a high proportion of macrophages and monocyte-like cells, which are known to play a role in immune responses and stress management. Cluster 2 has a high hybrid similarity score of 0.8389 and is characterized by the presence of genes involved in stress response, such as HSPA1A, HSPA1B, and SELENOP. These genes are also typically upregulated in response to cellular stress, suggesting that this cluster may be involved in the epithelial stress response. The cluster also contains a high proportion of macrophages and monocyte-like cells, similar to Cluster 3. Cluster 6 has a moderate hybrid similarity score of 0.8367 and is characterized by the presence of genes involved in stress response, such as HBA1, HBA2, and HBZ. These genes are typically upregulated in response to cellular stress, suggesting that this cluster may be involved in the epithelial stress response. The cluster also contains a high proportion of monocytes and macrophages, which are known to play a role in immune responses and stress management. Cluster 8 has a moderate hybrid similarity score of 0.8367 and is characterized by the presence of genes involved in stress response, such as RPS26, C1QC, and DAB2. These genes are typically upregulated in response to cellular stress, suggesting that this cluster may be involved in the epithelial stress response. The cluster also contains a high proportion of monocytes and macrophages, similar to Cluster 6. Cluster 13 has a moderate hybrid similarity score of 0.8349 and is characterized by the presence of genes involved in stress response, such as TTR, CD74, and PLXDC2. These genes are typically upregulated in response to cellular stress, suggesting that this cluster may be involved in the epithelial stress response. The cluster also contains a high proportion of ChP microglia and dendritic cells, which are known to play a role in immune responses and stress management. In summary, the clusters related to the epithelial stress response are characterized by the presence of genes involved in stress response and a high proportion of immune cells, such as macrophages and monocytes. These clusters may be involved in the epithelial stress response, but further analysis is needed to confirm this hypothesis.

scGPT: Epithelial stress response.

Based on the provided dataset evidence, I will analyze the clusters related to the epithelial stress response. Cluster 6 has a high hybrid similarity (0.9997496) and is primarily composed of epithelial cells (0.9988512). The cluster is enriched with genes such as SLPI, SCGB3A1, and KRT19, which are typically associated with epithelial cells. This cluster is likely related to the epithelial stress response, as it contains genes that are involved in maintaining epithelial integrity and function. Cluster 1 has a high hybrid similarity (0.9996884) and is also primarily composed of epithelial cells (0.9995779). This cluster is enriched with genes such as KRT17, MIR205HG, and JUNB, which are involved in cell proliferation, differentiation, and stress response. This cluster is likely related to the epithelial stress response, as it contains genes that are involved in regulating epithelial cell behavior. Cluster 17 has a high hybrid similarity (0.9996237) and is primarily composed of immune cells (0.9395629). However, this cluster also contains a significant proportion of epithelial cells (0.060437). The cluster is enriched with genes such as HLA-DRA, HLA-DPB1, and CD74, which are involved in immune cell function and antigen presentation. This cluster is likely related to the epithelial stress response, as it contains genes that are involved in the immune response to epithelial cells. Cluster 3 has a high hybrid similarity (0.9996114) and is primarily composed of immune cells (0.9973902). However, this cluster also contains a significant proportion of epithelial cells (0.0026098). The cluster is enriched with genes such as CCL5, PTPRC, and SYTL3, which are involved in immune cell function and cell signaling. This cluster is likely related to the epithelial stress response, as it contains genes that are involved in regulating immune cell behavior in response to epithelial cells. In summary, clusters 6, 1, 17, and 3 are likely related to the epithelial stress response, as they contain genes that are involved in maintaining epithelial integrity, regulating epithelial cell behavior, and responding to epithelial cells. However, it is essential to note that these clusters are not exclusively composed of epithelial cells and contain a significant proportion of immune cells.

Semantic: Oxidative stress signaling in epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to oxidative stress signaling in epithelial cells. Cluster 3 has a semantic similarity of 0.27004897594451904 and contains 667 cells. The cluster is primarily composed of macrophages and monocyte-like cells. The gene evidence shows that HSPA1B, HSPA1A, and SLC9A9 are highly expressed in this cluster, which are involved in heat shock protein activity and sodium transport, respectively. These genes may be related to cellular stress response and ion homeostasis. Cluster 15 has a semantic similarity of 0.2658652067184448 and contains 146 cells. The cluster is primarily composed of megakaryocytes and platelets. The gene evidence shows that CNST, LTBP1, and LIMS1 are highly expressed in this cluster, which are involved in cell adhesion and cytoskeleton organization. These genes may be related to cellular structure and function. Cluster 2 has a semantic similarity of 0.2619014382362366 and contains 958 cells. The cluster is primarily composed of macrophages and monocyte-like cells. The gene evidence shows that RPS4Y1, HSPA1A, and HSPA1B are highly expressed in this cluster, which are involved in ribosomal protein activity and heat shock protein activity, respectively. These genes may be related to cellular stress response and protein synthesis. Cluster 11 has a semantic similarity of 0.23533692955970764 and contains 270 cells. The cluster is primarily composed of microglia and dendritic cells. The gene evidence shows that PLXDC2, S100A11, and FRMD4A are highly expressed in this cluster, which are involved in cell adhesion and cytoskeleton organization. These genes may be related to cellular structure and function. Cluster 16 has a semantic similarity of 0.23141954839229584 and contains 141 cells. The cluster is primarily composed of macrophages and monocyte-like cells. The gene evidence shows that GPNMB, NEAT1, and SLC2A1 are highly expressed in this cluster, which are involved in cell adhesion and glucose transport, respectively. These genes may be related to cellular structure and function. In summary, the clusters related to oxidative stress signaling in epithelial cells are primarily composed of macrophages and monocyte-like cells. The highly expressed genes in these clusters are involved in cellular stress response, ion homeostasis, cell adhesion, and cytoskeleton organization. However, it is essential to note that the provided dataset evidence does not explicitly mention oxidative stress signaling, and the relevance of these clusters to epithelial cells is not directly established. Further

analysis and experimental validation are required to confirm the biological significance of these findings.

Hybrid: Oxidative stress signaling in epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to oxidative stress signaling in epithelial cells. Cluster 3 has a high hybrid similarity score (0.853746) and is characterized by a high expression of heat shock proteins (HSPs) such as HSPA1B, HSPA1A, and HSPB1. These proteins are known to play a role in protecting cells against oxidative stress. Additionally, the cluster has a high percentage of cells expressing genes involved in iron metabolism, which is relevant to oxidative stress. Cluster 15 has a moderate hybrid similarity score (0.852924) and is characterized by a high expression of genes involved in cell adhesion and migration, such as CNST and LIMS1. While these genes are not directly related to oxidative stress, they may be involved in the cellular response to oxidative stress. Cluster 2 has a high hybrid similarity score (0.851491) and is characterized by a high expression of HSPs, including HSPA1A, HSPA1B, and HSPB1. This cluster also has a high percentage of cells expressing genes involved in iron metabolism, similar to Cluster 3. Cluster 11 has a moderate hybrid similarity score (0.846779) and is characterized by a high expression of genes involved in immune responses, such as TREM2 and ALOX5AP. While these genes are not directly related to oxidative stress, they may be involved in the cellular response to oxidative stress. Cluster 16 has a moderate hybrid similarity score (0.846182) and is characterized by a high expression of genes involved in cell survival and proliferation, such as GPNMB and NEAT1. While these genes are not directly related to oxidative stress, they may be involved in the cellular response to oxidative stress. In summary, clusters 3, 2, and 16 have a high expression of heat shock proteins and genes involved in iron metabolism, which are relevant to oxidative stress signaling in epithelial cells.

scGPT: Oxidative stress signaling in epithelial cells.

Based on the provided dataset evidence, we can analyze the relevance of each cluster to oxidative stress signaling in epithelial cells. Cluster 6 has a high hybrid similarity score (0.9997572302818298) and is predominantly composed of epithelial cells (0.9988512349224583). This cluster is likely relevant to oxidative stress signaling in epithelial cells, as it contains genes such as SLPI, SCGB3A1, and KRT19, which are involved in cellular defense mechanisms and epithelial cell function. Cluster 1 has a high hybrid similarity score (0.9996640086174011) and is also predominantly composed of epithelial cells (0.9995779165963194). This cluster is likely relevant to oxidative stress signaling in epithelial cells, as it contains genes such as KRT17, FOS, and HSPB1, which are involved in cellular stress responses and epithelial cell function. Cluster 3 has a high hybrid similarity score (0.9996609687805176) and is predominantly composed of immune cells (0.9973902364138778). However, this cluster also contains a small proportion of epithelial cells (0.0026097635861221983), which may be relevant to oxidative stress signaling in epithelial cells. This cluster contains genes such as CCL5, PTPRC, and S100A4, which are involved in immune cell function and may be relevant to oxidative stress responses. Cluster 15 has a high hybrid similarity score (0.9996118545532227) and is predominantly composed of immune cells (0.9982046678635548). However, this cluster also contains a small proportion of epithelial cells (0.0017953321364452424), which may be relevant to oxidative stress signaling in epithelial cells. This cluster contains genes such as PTPRC, CCL5, and SRGN, which are involved in immune cell function and may be relevant to oxidative stress responses. Cluster 17 has a high hybrid similarity score (0.9995568990707397) and is predominantly composed of immune cells (0.9395629939562994). However, this cluster also contains a small proportion of epithelial cells (0.06043700604370061), which may be relevant to oxidative stress signaling in epithelial cells. This cluster contains genes such as HLA-DRA, HLA-DPB1, and CD74, which are involved in immune cell function and may be relevant to oxidative stress responses. In summary, clusters 6 and 1 are likely the most relevant to oxidative stress signaling in epithelial cells, as they contain a high proportion of epithelial cells

and genes involved in cellular defense mechanisms and epithelial cell function. Clusters 3, 15, and 17 may also be relevant, as they contain a small proportion of epithelial cells and genes involved in immune cell function and oxidative stress responses.

Semantic: Immune-like signaling in epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to immune-like signaling in epithelial cells. Cluster 15 has a semantic similarity of 0.1933 and contains 146 cells. The metadata indicates that this cluster is primarily composed of megakaryocytes and platelets, with a development stage ontology term ID of HsapDv:0000047. This suggests that the cells in this cluster are likely involved in blood cell development. However, there is no clear indication of immune-like signaling in epithelial cells. Cluster 1 has a semantic similarity of 0.1870 and contains 1246 cells. The metadata shows that this cluster is primarily composed of macrophages, with a development stage ontology term ID of HsapDv:0000027. The gene evidence includes genes such as C1QA, DAB2, and FCGRT, which are involved in immune responses. However, there is no clear indication that these cells are epithelial in origin. Cluster 11 has a semantic similarity of 0.1736 and contains 270 cells. The metadata indicates that this cluster is primarily composed of microglia and dendritic cells, with a development stage ontology term ID of HsapDv:0000027. The gene evidence includes genes such as PLXDC2, S100A11, and TREM2, which are involved in immune responses. However, there is no clear indication that these cells are epithelial in origin. Cluster 2 has a semantic similarity of 0.1726 and contains 958 cells. The metadata shows that this cluster is primarily composed of macrophages-iron metabolism cells, with a development stage ontology term ID of HsapDv:0000050. The gene evidence includes genes such as RPS4Y1, HSPA1A, and HSPA1B, which are involved in immune responses. However, there is no clear indication that these cells are epithelial in origin. Cluster 16 has a semantic similarity of 0.1582 and contains 141 cells. The metadata indicates that this cluster is primarily composed of macrophages-iron metabolism cells, with a development stage ontology term ID of HsapDv:0000050. The gene evidence includes genes such as GPNMB, NEAT1, and SLC2A1, which are involved in immune responses. However, there is no clear indication that these cells are epithelial in origin. In summary, while the dataset evidence suggests that there are immune-like signaling pathways present in some of the clusters, there is no clear indication that these cells are epithelial in origin. The clusters appear to be primarily composed of immune cells such as macrophages and microglia.

Hybrid: Immune-like signaling in epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to immune-like signaling in epithelial cells. Cluster 15 has a high hybrid similarity score (0.8383875489234924) and is characterized by a high percentage of cells classified as "Myeloid" (1.0). However, the gene expression evidence does not explicitly indicate immune-like signaling in epithelial cells. The genes present in this cluster, such as CNST, LTBP1, and LIMS1, are not typically associated with immune-like signaling. Cluster 1 has a high hybrid similarity score (0.8371297121047974) and is also characterized by a high percentage of cells classified as "Myeloid" (1.0). The gene expression evidence in this cluster includes genes such as LGMN, C1QA, and FCGRT, which are involved in immune responses. However, the cluster is not specifically related to epithelial cells. Cluster 11 has a moderate hybrid similarity score (0.834468424320221) and is characterized by a high percentage of cells classified as "Microglia" (0.7148148148148148). The gene expression evidence in this cluster includes genes such as PLXDC2, S100A11, and TREM2, which are involved in immune responses. However, the cluster is not specifically related to epithelial cells. Cluster 2 has a moderate hybrid similarity score (0.8336092829704285) and is characterized by a high percentage of cells classified as "Macrophages-iron metabolism" (0.8507306889352818). The gene expression evidence in this cluster includes genes such as RPS4Y1, HSPA1A, and HSPA1B, which are involved in immune responses. However, the cluster is not specifically related to epithelial cells. Cluster 16 has a moderate hybrid similar-

ity score (0.8315550684928894) and is characterized by a high percentage of cells classified as "Macrophages-iron metabolism" (0.9645390070921985). The gene expression evidence in this cluster includes genes such as GPNMB, NEAT1, and KLF6, which are involved in immune responses. However, the cluster is not specifically related to epithelial cells. In summary, while there are clusters with high hybrid similarity scores and gene expression evidence related to immune responses, none of them are specifically related to immune-like signaling in epithelial cells. The provided dataset evidence does not provide clear evidence for immune-like signaling in epithelial cells.

scGPT: Immune-like signaling in epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to immune-like signaling in epithelial cells. Cluster 6 (n_cells = 5223) is the most relevant to immune-like signaling in epithelial cells, with a high hybrid similarity of 0.9997513294219971. This cluster is predominantly epithelial (0.9988512349224583) and has a low immune cell fraction (0.0009573042312847023). The genes present in this cluster are involved in various cellular processes, including: - SLPI (Secretory Leukocyte Peptide 1): involved in antimicrobial defense and immune regulation. - SCGB3A1 (Small Cysteine-Rich Protein 3A1): involved in immune response and epithelial cell function. - KRT19 (Keratin 19): a marker for epithelial cells. Cluster 1 (n_cells = 11846) is also relevant, with a high hybrid similarity of 0.9996824264526367. This cluster is predominantly epithelial (0.9995779165963194) and has a low immune cell fraction (0.00033766672294445384). The genes present in this cluster are involved in various cellular processes, including: - KRT17 (Keratin 17): a marker for epithelial cells. - CD9 (CD9 Molecule): involved in cell adhesion and immune response. Cluster 15 (n_cells = 2228) is an immune-like cluster, with a high hybrid similarity of 0.999625563621521. This cluster is predominantly immune (0.9982046678635548) and has a low epithelial cell fraction (0.0017953321364452424). The genes present in this cluster are involved in various immune processes, including: - PTPRC (Protein Tyrosine Phosphatase, Receptor Type C): involved in immune cell signaling. - CCL5 (Chemokine (C-C Motif) Ligand 5): involved in immune cell recruitment. Cluster 17 (n_cells = 2151) is also an immune-like cluster, with a high hybrid similarity of 0.9995944499969482. This cluster is predominantly immune (0.9395629939562994) and has a low epithelial cell fraction (0.06043700604370061). The genes present in this cluster are involved in various immune processes, including: - HLA-DRA (Major Histocompatibility Complex, Class II, DR alpha): involved in immune cell recognition. - CD74 (CD74 Molecule): involved in immune cell signaling. Cluster 3 (n_cells = 6514) is an immune-like cluster, with a high hybrid similarity of 0.9995903372764587. This cluster is predominantly immune (0.9973902364138778) and has a low epithelial cell fraction (0.0026097635861221983). The genes present in this cluster are involved in various immune processes, including: - CCL5 (Chemokine (C-C Motif) Ligand 5): involved in immune cell recruitment. - PTPRC (Protein Tyrosine Phosphatase, Receptor Type C): involved in immune cell signaling. In summary, clusters 6, 1, 15, 17, and 3 are relevant to immune-like signaling in epithelial cells, with genes involved in various cellular processes, including immune response, epithelial cell function, and immune cell signaling.

Semantic: Interferon signaling in airway epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to interferon signaling in airway epithelial cells. Cluster 12 has a semantic similarity of 0.2156757116317749. The metadata indicates that this cluster consists of 258 cells, primarily annotated as Myeloid cells with a subclass of Macrophages. However, there is no explicit evidence of interferon signaling or airway epithelial cells in this cluster. Cluster 15 has a semantic similarity of 0.20156925916671753. The metadata shows that this cluster consists of 146 cells, primarily annotated as Myeloid cells with a subclass of Megakaryocytes. There is no explicit evidence of interferon signaling or airway epithelial cells in this cluster. Cluster 14 has a semantic similarity of 0.192769393324852. The metadata in-

icates that this cluster consists of 157 cells, primarily annotated as Myeloid cells with a subclass of Macrophages Monocyte-like. However, there is no explicit evidence of interferon signaling or airway epithelial cells in this cluster. Cluster 13 has a semantic similarity of 0.17960962653160095. The metadata shows that this cluster consists of 165 cells, primarily annotated as Myeloid cells with a subclass of ChP Microglia. There is no explicit evidence of interferon signaling or airway epithelial cells in this cluster. Cluster 8 has a semantic similarity of 0.17816029489040375. The metadata indicates that this cluster consists of 342 cells, primarily annotated as Myeloid cells with a subclass of Monocytes. However, there is no explicit evidence of interferon signaling or airway epithelial cells in this cluster. Unfortunately, none of the clusters provided in the dataset evidence explicitly relate to interferon signaling in airway epithelial cells. The dataset does not contain any information about airway epithelial cells or interferon signaling pathways. Therefore, I cannot provide a biological explanation for cluster relevance based on the provided evidence.

Hybrid: Interferon signaling in airway epithelial cells.

Based on the provided dataset evidence, I will analyze the relevance of each cluster to interferon signaling in airway epithelial cells. Cluster 12 has a high hybrid similarity score (0.8427445292472839) and is primarily composed of macrophages (Subclass: Macrophages, 86.04651162790697) the presence of genes such as TUBB, TUBA1B, and H2AZ1, which are involved in cellular processes, suggests a general cellular response. Cluster 15 has a moderate hybrid similarity score (0.8400502800941467) and is primarily composed of megakaryocytes and platelets (Subclass: Megakaryocytes, 68.4931506849315) airway epithelial cells. Cluster 14 has a moderate hybrid similarity score (0.8379185795783997) and is primarily composed of macrophages monocyte-like cells (Subclass: Macrophages Monocyte-like, 63.05732484076433) the presence of genes such as TUBA1B, STMN1, and H2AZ1, which are involved in cellular processes, suggests a general cellular response. Cluster 8 has a moderate hybrid similarity score (0.8353995084762573) and is primarily composed of monocytes and macrophages (Subclass: Monocytes, 82.16374269005848) the presence of genes such as RPS26, C1QC, and RPL39, which are involved in cellular processes, suggests a general cellular response. Cluster 13 has a moderate hybrid similarity score (0.835165798664093) and is primarily composed of microglia (Subclass: ChP Microglia, 99.393939393939) the presence of genes such as TTR, CD74, and PLXDC2, which are involved in cellular processes, suggests a general cellular response. In summary, none of the clusters have direct evidence of interferon signaling in airway epithelial cells. However, clusters 12, 14, 8, and 13 show general cellular responses, suggesting that these cells may be involved in responding to interferon signaling in airway epithelial cells.

scGPT: Interferon signaling in airway epithelial cells.

Based on the provided dataset evidence, I will analyze the clusters related to interferon signaling in airway epithelial cells. Cluster 6 has a high hybrid similarity of 0.9997431635856628 and is primarily composed of airway epithelial cells (0.9988512349224583). The cluster is predominantly found in the control group (Ctrl) with a proportion of 0.7537813517135745. This cluster may represent a baseline or normal state of airway epithelial cells. Cluster 15 has a high hybrid similarity of 0.9996976852416992 and is primarily composed of immune cells (0.9982046678635548). The cluster is predominantly found in the control group (Ctrl) with a proportion of 0.8096947935368043. This cluster may represent a normal state of immune cells. Cluster 1 has a high hybrid similarity of 0.9996824860572815 and is primarily composed of airway epithelial cells (0.9995779165963194). The cluster is predominantly found in the control group (Ctrl) with a proportion of 0.975519162586527. This cluster may represent a normal state of airway epithelial cells. Cluster 3 has a high hybrid similarity of 0.9996627569198608 and is primarily composed of immune cells (0.9973902364138778). The cluster is predominantly found in the control group (Ctrl) with a proportion of 0.9886398526251151. This cluster may represent a normal state of immune cells. Cluster 17 has a high hybrid similarity of

0.9996330738067627 and is primarily composed of immune cells (0.9395629939562994). The cluster is predominantly found in the control group (Ctrl) with a proportion of 0.891213389121339. This cluster may represent a normal state of immune cells. In terms of interferon signaling, the clusters do not explicitly mention interferon-related genes. However, some genes mentioned in the evidence, such as CCL5, PTPRC, and SRGN, are involved in immune responses and may be related to interferon signaling indirectly. In conclusion, the clusters primarily represent normal states of airway epithelial cells and immune cells. While there is no direct evidence of interferon signaling, some genes mentioned in the evidence may be related to immune responses and potentially interferon signaling.

Sematic: Antigen presentation in epithelial. clusters.

Based on the provided dataset evidence, I will analyze the relevance of each cluster to antigen presentation in epithelial cells. Cluster 18 has a semantic similarity of 0.21968895196914673 and is primarily composed of mast cells (0.7888888888888889). While mast cells are involved in immune responses, there is no direct evidence of antigen presentation in epithelial cells in this cluster. Cluster 1 has a semantic similarity of 0.2105187475681305 and is primarily composed of macrophages (0.7407704654895666). Macrophages are known to play a crucial role in antigen presentation, particularly in the context of epithelial cells. This cluster is likely relevant to antigen presentation in epithelial cells. Cluster 15 has a semantic similarity of 0.20208333432674408 and is primarily composed of megakaryocytes (0.684931506849315) and platelets (0.3082191780821918). While megakaryocytes and platelets are involved in blood coagulation and immune responses, there is no direct evidence of antigen presentation in epithelial cells in this cluster. Cluster 11 has a semantic similarity of 0.18962669372558594 and is primarily composed of microglia (0.7148148148148148) and dendritic cells (0.2111111111111111). Microglia and dendritic cells are involved in antigen presentation, particularly in the context of the central nervous system. This cluster is likely relevant to antigen presentation in epithelial cells, particularly in the context of the meninges. Cluster 20 has a semantic similarity of 0.16995032131671906 and is primarily composed of ILC precursors (1.0). ILC precursors are involved in immune responses, but there is no direct evidence of antigen presentation in epithelial cells in this cluster. In summary, clusters 1 and 11 are likely relevant to antigen presentation in epithelial cells, particularly in the context of macrophages and microglia/dendritic cells, respectively.

Hybrid: Antigen presentation in epithelial. clusters.

Based on the provided dataset evidence, we can observe that there are several clusters related to antigen presentation in epithelial cells. However, it's essential to note that none of the clusters are explicitly labeled as "epithelial" or "antigen presentation." Instead, we can infer their relevance based on the gene expression profiles and metadata. Cluster 18 has a high hybrid similarity score (0.843683660030365) and is primarily composed of myeloid cells, with a significant proportion of mast cells (0.7888888888888889). While there are no specific genes related to antigen presentation in this cluster, the presence of mast cells suggests a potential role in immune responses. Cluster 1 is also predominantly myeloid, with a high proportion of macrophages (0.7407704654895666). This cluster has a high hybrid similarity score (0.8418456315994263) and contains several genes related to antigen presentation, including LGMN, C1QA, and FCGRT. These genes are often associated with the processing and presentation of antigens to T-cells. Cluster 15 is primarily composed of megakaryocytes and platelets (0.684931506849315 and 0.3082191780821918, respectively). While this cluster has a moderate hybrid similarity score (0.8401198983192444), it does not contain any genes explicitly related to antigen presentation. Cluster 11 is composed of microglia and dendritic cells (0.7148148148148148 and 0.2111111111111111, respectively). This cluster has a moderate hybrid similarity score (0.8376960158348083) and contains several genes related to antigen presentation, including PLXDC2, S100A11, and TREM2. Cluster 20 is primarily composed of ILC precursors (1.0) and has a moderate hybrid similarity score (0.8337268829345703). While this cluster does not contain any genes explicitly related to

antigen presentation, it is worth noting that ILCs can play a role in immune responses and may interact with antigen-presenting cells. In summary, while none of the clusters are explicitly labeled as "epithelial" or "antigen presentation," clusters 1 and 11 contain genes related to antigen presentation, and cluster 18 has a significant proportion of mast cells, which may suggest a potential role in immune responses. However, further analysis and validation are necessary to confirm these findings.

scGPT: Antigen presentation in epithelial.

Based on the provided dataset evidence, we can analyze the relevance of each cluster to antigen presentation in epithelial cells. Cluster 1 (n_cells = 11846) is predominantly epithelial (Epithelial: 0.9995779165963194) with a small fraction of immune cells (Immune: 0.00033766672294445384). This cluster does not contain any genes explicitly related to antigen presentation. Cluster 6 (n_cells = 5223) is also predominantly epithelial (Epithelial: 0.9988512349224583) with a small fraction of immune cells (Immune: 0.0009573042312847023). This cluster contains genes related to the epithelial barrier, such as SLPI (pct_in: 0.9923415661497224) and SCGB3A1 (pct_in: 0.9860233582232434), but not directly related to antigen presentation. Cluster 15 (n_cells = 2228) is predominantly immune (Immune: 0.9982046678635548) with a small fraction of epithelial cells (Epithelial: 0.0017953321364452424). This cluster contains genes related to immune cells, such as PTPRC (pct_in: 0.9169658886894075) and CCL5 (pct_in: 0.8729802513464991), but not directly related to antigen presentation in epithelial cells. Cluster 17 (n_cells = 2151) is predominantly immune (Immune: 0.9395629939562994) with a small fraction of epithelial cells (Epithelial: 0.06043700604370061). This cluster contains genes related to antigen presentation, such as HLA-DRA (pct_in: 0.99721059972106) and HLA-DPB1 (pct_in: 0.9916317991631799), which are major histocompatibility complex (MHC) class II molecules involved in antigen presentation. Cluster 8 (n_cells = 3964) is predominantly immune (Immune: 0.9987386478304743) with a small fraction of epithelial cells (Epithelial: 0.0012613521695257316). This cluster contains genes related to immune cells, such as PTPRC (pct_in: 0.9631685166498486) and FYN (pct_in: 0.9492936427850656), but not directly related to antigen presentation in epithelial cells. In conclusion, Cluster 17 is the most relevant to antigen presentation in epithelial cells, containing genes related to MHC class II molecules involved in antigen presentation.

Semantic: MHC-II expression in epithelial cells

Based on the provided dataset evidence, I will analyze the relevance of each cluster to MHC-II expression in epithelial cells. Cluster 2 (semantic similarity: 0.32849860191345215) is primarily composed of macrophages, with a high percentage of cells classified as "Macrophages-iron metabolism" (0.8507306889352818). This cluster does not explicitly mention MHC-II expression or epithelial cells. However, macrophages are known to interact with epithelial cells in various contexts, including immune responses. Cluster 14 (semantic similarity: 0.3198525607585907) is also primarily composed of macrophages, with a high percentage of cells classified as "Macrophages Monocyte-like" (0.6305732484076433). This cluster does not explicitly mention MHC-II expression or epithelial cells. Like Cluster 2, macrophages in this cluster may interact with epithelial cells. Cluster 18 (semantic similarity: 0.3161256015300751) is primarily composed of mast cells, with a high percentage of cells classified as "Mast cells" (0.7888888888888889). This cluster does not explicitly mention MHC-II expression or epithelial cells. Mast cells are known to interact with epithelial cells in various contexts, including allergic responses. Cluster 1 (semantic similarity: 0.31358852982521057) is primarily composed of macrophages, with a high percentage of cells classified as "Macrophages" (0.7407704654895666). This cluster does not explicitly mention MHC-II expression or epithelial cells. Like Clusters 2 and 14, macrophages in this cluster may interact with epithelial cells. Cluster 11 (semantic similarity: 0.30888211727142334) is primarily composed of microglia, with a high percentage of cells classified as "Microglia" (0.7148148148148148). This cluster does not explicitly men-

tion MHC-II expression or epithelial cells. Microglia are known to interact with epithelial cells in various contexts, including immune responses. In summary, none of the clusters explicitly mention MHC-II expression or epithelial cells. However, macrophages and microglia in these clusters may interact with epithelial cells in various contexts. The provided dataset evidence does not provide direct evidence of MHC-II expression in epithelial cells.

Hybrid: MHC-II expression in epithelial cells

Based on the provided dataset evidence, I will analyze the clusters related to MHC-II expression in epithelial cells. Cluster 2 has a high hybrid similarity of 0.8644 and is associated with macrophages, particularly macrophages-iron metabolism (0.85). However, there is no explicit evidence of MHC-II expression in this cluster. Cluster 14 has a high hybrid similarity of 0.8619 and is also associated with macrophages, including macrophages-monocyte-like (0.63) and macrophages-iron metabolism (0.32). Notably, this cluster contains genes related to antigen presentation, such as H2AZ1 (pct.in = 1.0) and TUBA1B (pct.in = 1.0). While these genes are involved in antigen presentation, they are not specific to MHC-II expression. Cluster 18 has a high hybrid similarity of 0.8619 and is associated with mast cells (0.79) and monocytes (0.01). This cluster contains genes related to immune response, such as S100A6 (pct.in = 0.89) and TIMP3 (pct.in = 0.89). However, there is no explicit evidence of MHC-II expression in this cluster. Cluster 1 has a high hybrid similarity of 0.8612 and is associated with macrophages, including macrophages (0.74) and monocytes (0.05). This cluster contains genes related to antigen presentation, such as C1QA (pct.in = 0.89) and CD14 (pct.in = 0.83). Notably, CD14 is a receptor for MHC-II molecules, suggesting a potential link to MHC-II expression. Cluster 11 has a high hybrid similarity of 0.8607 and is associated with microglia (0.71) and dendritic cells (0.21). This cluster contains genes related to immune response, such as PLXDC2 (pct.in = 0.96) and S100A11 (pct.in = 0.87). However, there is no explicit evidence of MHC-II expression in this cluster. In summary, while there are clusters associated with macrophages and antigen presentation, there is no explicit evidence of MHC-II expression in epithelial cells in the provided dataset. However, Cluster 1 contains genes related to antigen presentation, including CD14, which is a receptor for MHC-II molecules, suggesting a potential link to MHC-II expression.

scGPT: MHC-II expression in epithelial cells

Based on the provided dataset evidence, we can observe that there are multiple clusters with varying levels of MHC-II expression. Cluster 6 has a high level of MHC-II expression (hybrid similarity: 0.9997583627700806) and is primarily composed of epithelial cells (Epithelial: 0.9988512349224583). This cluster is likely to be relevant for studying MHC-II expression in epithelial cells. Cluster 3 has a moderate level of MHC-II expression (hybrid similarity: 0.9996510148048401) and is primarily composed of immune cells (Immune: 0.9973902364138778). This cluster is likely to be relevant for studying MHC-II expression in immune cells. Cluster 17 has a high level of MHC-II expression (hybrid similarity: 0.9995047450065613) and is primarily composed of immune cells (Immune: 0.9395629939562994). This cluster is likely to be relevant for studying MHC-II expression in immune cells. The presence of HLA-DRA, HLA-DPB1, CD74, HLA-DQA1, and other genes associated with MHC-II expression in Cluster 17 suggests that this cluster is relevant for studying MHC-II expression in immune cells. In summary, Clusters 6, 3, and 17 are relevant for studying MHC-II expression in epithelial and immune cells, respectively.

Semantic: Inter-epithelial immune signaling.

Based on the provided dataset evidence, I will analyze the relevance of each cluster to inter-epithelial immune signaling. Cluster 11 has a semantic similarity of 0.16916358470916748 and is primarily composed of myeloid cells, including microglia and dendritic cells. The presence of genes such as PLXDC2, S100A11, and TREM2 suggests involvement in immune responses, particularly in the context of the central nervous system. Cluster 15 has

a semantic similarity of 0.16838136315345764 and is also primarily composed of myeloid cells, including megakaryocytes and platelets. The presence of genes such as CNST, LIMS1, and RGS18 suggests involvement in immune responses, particularly in the context of blood cell development and function. Cluster 13 has a semantic similarity of 0.1672395020723343 and is primarily composed of myeloid cells, including ChP microglia and dendritic cells. The presence of genes such as PLXDC2, S100A11, and CD74 suggests involvement in immune responses, particularly in the context of the central nervous system. Cluster 18 has a semantic similarity of 0.167183518409729 and is primarily composed of myeloid cells, including mast cells and monocytes. The presence of genes such as ARHGAP15, GATA2, and S100A6 suggests involvement in immune responses, particularly in the context of allergic reactions and inflammation. Cluster 8 has a semantic similarity of 0.16518470644950867 and is primarily composed of myeloid cells, including monocytes and macrophages. The presence of genes such as RPS26, C1QC, and DAB2 suggests involvement in immune responses, particularly in the context of innate immunity and inflammation. In summary, clusters 11, 15, 13, 18, and 8 are all relevant to inter-epithelial immune signaling, with a focus on myeloid cell development and function, immune responses in the central nervous system, and inflammation. However, it is essential to note that these clusters are not exclusively involved in inter-epithelial immune signaling, and further analysis would be required to determine the specific mechanisms and pathways involved.

Hybrid: Inter-epithelial immune signaling.

Based on the provided dataset evidence, I will analyze the clusters related to inter-epithelial immune signaling. Cluster 11 has a high hybrid similarity of 0.8335869312286377. The cluster is primarily composed of myeloid cells, with a high proportion of microglia (71.48) contains genes related to immune signaling, such as PLXDC2, SLC8A1, and S100A11. These genes are involved in various immune-related processes, including antigen presentation and cytokine signaling. Cluster 15 has a high hybrid similarity of 0.8334002494812012. The cluster is also primarily composed of myeloid cells, with a high proportion of megakaryocytes (68.49) genes related to immune signaling, such as CNST, LTBP1, and LIMS1. These genes are involved in various immune-related processes, including platelet production and cytokine signaling. Cluster 18 has a high hybrid similarity of 0.8331824541091919. The cluster is primarily composed of myeloid cells, with a high proportion of mast cells (78.89) signaling, such as ARHGAP15, GATA2, and S100A6. These genes are involved in various immune-related processes, including mast cell activation and cytokine signaling. Cluster 8 has a high hybrid similarity of 0.8328137993812561. The cluster is primarily composed of myeloid cells, with a high proportion of monocytes (82.16) RPS26, C1QC, and DAB2. These genes are involved in various immune-related processes, including antigen presentation and cytokine signaling. Cluster 13 has a high hybrid similarity of 0.83269202709198. The cluster is primarily composed of myeloid cells, with a high proportion of ChP microglia (99.39) PLXDC2. These genes are involved in various immune-related processes, including antigen presentation and cytokine signaling. In summary, the clusters related to inter-epithelial immune signaling are primarily composed of myeloid cells, with a high proportion of microglia, megakaryocytes, mast cells, monocytes, and ChP microglia. The clusters contain genes related to immune signaling, including PLXDC2, SLC8A1, S100A11, CNST, LTBP1, LIMS1, ARHGAP15, GATA2, S100A6, RPS26, C1QC, DAB2, TTR, CD74, and PLXDC2. These genes are involved in various immune-related processes, including antigen presentation, cytokine signaling, and mast cell activation.

scGPT: Inter-epithelial immune signaling.

Based on the provided dataset evidence, I will analyze the clusters related to inter-epithelial immune signaling. Cluster 6 (hybrid_similarity: 0.9997289776802063) is primarily composed of epithelial cells (0.9988512349224583) with a minor presence of immune cells (0.0009573042312847023). This cluster is likely involved in the mainte-

nance of epithelial barrier function and may play a role in the regulation of immune responses at the epithelial interface. Cluster 1 (hybrid_similarity: 0.9996324777603149) is predominantly composed of epithelial cells (0.9995779165963194) with a minor presence of immune cells (0.00033766672294445384). This cluster is likely involved in the regulation of epithelial cell function and may play a role in the maintenance of epithelial homeostasis. Cluster 3 (hybrid_similarity: 0.9996134638786316) is primarily composed of immune cells (0.9973902364138778) with a minor presence of epithelial cells (0.0026097635861221983). This cluster is likely involved in the regulation of immune responses and may play a role in the coordination of immune cell function at the epithelial interface. Cluster 15 (hybrid_similarity: 0.9995874166488647) is primarily composed of immune cells (0.9982046678635548) with a minor presence of epithelial cells (0.0017953321364452424). This cluster is likely involved in the regulation of immune responses and may play a role in the coordination of immune cell function at the epithelial interface. Cluster 17 (hybrid_similarity: 0.9995312094688416) is primarily composed of immune cells (0.9395629939562994) with a minor presence of epithelial cells (0.06043700604370061). This cluster is likely involved in the regulation of immune responses and may play a role in the coordination of immune cell function at the epithelial interface. The genes present in the dataset evidence are: - Cluster 6: SLPI, SCGB3A1, SCGB1A1, KRT19, WFDC2, BPIFB1, MSMB, PIGR, CHST9, LCN2 - Cluster 1: KRT17, MIR205HG, CD9, JUNB, ZSWIM6, FOS, PRSS23, KRT15, HSPB1, S100A2 - Cluster 3: CCL5, PTPRC, SYTL3, FYN, S100A4, CXCR4, SRGN, SAMSN1, SIK3, STAT4 - Cluster 15: PTPRC, CCL5, SRGN, FYN, NKG7, IL32, SYTL3, SKAP1, ARHGAP15, AOA - Cluster 17: HLA-DRA, HLA-DPB1, CD74, HLA-DQA1, AIF1, TYROBP, LYZ, LGALS1, FCER1G, SRGN These genes are likely involved in the regulation of epithelial barrier function, immune responses, and the coordination of immune cell function at the epithelial interface. However, further analysis is required to determine their specific roles in inter-epithelial immune signaling.

Semantic: Cytokine signaling within epithelial.

Based on the provided dataset evidence, we can observe that there are four clusters related to cytokine signaling within epithelial cells. Cluster 15 has a semantic similarity of 0.24685706198215485 and is characterized by the presence of myeloid cells. However, there is no explicit evidence of cytokine signaling within epithelial cells in this cluster. Cluster 20 has a semantic similarity of 0.2295704334974289 and is primarily composed of lymphoid cells, specifically ILC precursors. The presence of genes such as CD52, TRBC1, and IL7R suggests an immune-related function, but it is not directly related to cytokine signaling within epithelial cells. Cluster 16 has a semantic similarity of 0.2273232787847519 and is primarily composed of myeloid cells, specifically macrophages. The presence of genes such as GPNMB, NEAT1, and KLF6 suggests an immune-related function, but it is not directly related to cytokine signaling within epithelial cells. Cluster 13 has a semantic similarity of 0.22030571103096008 and is primarily composed of myeloid cells, specifically microglia. The presence of genes such as TTR, CD74, and PLXDC2 suggests an immune-related function, but it is not directly related to cytokine signaling within epithelial cells. Cluster 11 has a semantic similarity of 0.22004388272762299 and is primarily composed of myeloid cells, specifically microglia and dendritic cells. The presence of genes such as PLXDC2, S100A11, and DOCK4 suggests an immune-related function, but it is not directly related to cytokine signaling within epithelial cells. In summary, none of the clusters directly related to cytokine signaling within epithelial cells. However, clusters 16, 13, and 11 are related to immune-related functions, which may be indirectly related to cytokine signaling.

Hybrid: Cytokine signaling within epithelial. compartments

Based on the provided dataset evidence, I will analyze the clusters related to cytokine signaling within epithelial cells. Cluster 15 has a high hybrid similarity score (0.849125) and is primarily composed of myeloid cells (ClassAnn: Myeloid = 1.0). Although there are no spe-

cific cytokine-related genes mentioned in the gene_evidence section, the presence of genes like LIMS1 (pct.in = 0.856164) and RGS18 (pct.in = 0.849315) might be related to cellular signaling processes. Cluster 20 has a moderate hybrid similarity score (0.845586) and is primarily composed of lymphoid cells (ClassAnn: Lymphoid = 1.0). The gene_evidence section includes cytokine-related genes like IL7R (pct.in 0.953846) and NFKBIA (pct.in = 0.953846), which are involved in immune responses and signaling pathways. Cluster 16 has a moderate hybrid similarity score (0.845372) and is primarily composed of myeloid cells (ClassAnn: Myeloid = 1.0). The gene_evidence section includes cytokine-related genes like KLF6 (pct.in = 0.964539) and FOS (pct.in = 0.957446), which are involved in cellular differentiation and immune responses. Cluster 11 has a moderate hybrid similarity score (0.843732) and is primarily composed of myeloid cells (ClassAnn: Myeloid = 1.0). The gene_evidence section includes cytokine-related genes like DOCK4 (pct.in = 0.848148) and TREM2 (pct.in = 0.837037), which are involved in immune responses and signaling pathways. Cluster 1 has a moderate hybrid similarity score (0.843601) and is primarily composed of myeloid cells (ClassAnn: Myeloid = 1.0). The gene_evidence section includes cytokine-related genes like LGMN (pct.in = 0.899678) and C1QA (pct.in = 0.892455), which are involved in immune responses and cellular signaling processes. In summary, clusters 20, 16, 11, and 1 show relevance to cytokine signaling within epithelial cells, primarily through the presence of cytokine-related genes like IL7R, NFKBIA, KLF6, FOS, DOCK4, TREM2, LGMN, and C1QA. However, it is essential to note that these genes are not exclusively cytokine-related and may have broader roles in cellular signaling processes.

scGPT: Cytokine signaling within epithelial.

Based on the provided dataset evidence, the clusters can be described as follows: Cluster 6: This cluster is predominantly composed of epithelial cells, with a high expression of genes such as SLPI, SCGB3A1, and KRT19. These genes are associated with the airway epithelium and are involved in the regulation of inflammation and immune responses. The cluster also shows a high expression of genes related to the epithelial barrier function, such as CHST9 and LCN2. Cluster 1: This cluster is also primarily composed of epithelial cells, with a high expression of genes such as KRT17, MIR205HG, and CD9. These genes are associated with the airway epithelium and are involved in the regulation of cell proliferation and differentiation. The cluster also shows a high expression of genes related to the immune response, such as FOS and PRSS23. Cluster 3: This cluster is predominantly composed of immune cells, with a high expression of genes such as CCL5, PTPRC, and FYN. These genes are associated with the regulation of immune responses and are involved in the activation of immune cells. The cluster also shows a high expression of genes related to the lymphoid lineage, such as S100A4 and CXCR4. Cluster 15: This cluster is also predominantly composed of immune cells, with a high expression of genes such as PTPRC, CCL5, and SRGN. These genes are associated with the regulation of immune responses and are involved in the activation of immune cells. The cluster also shows a high expression of genes related to the myeloid lineage, such as SYTL3 and SKAP1. Cluster 17: This cluster is predominantly composed of immune cells, with a high expression of genes such as HLA-DRA, HLA-DPB1, and CD74. These genes are associated with the regulation of immune responses and are involved in the presentation of antigens to T-cells. The cluster also shows a high expression of genes related to the myeloid lineage, such as TYROBP and LYZ. In summary, the clusters can be broadly categorized into epithelial and immune cell populations, with distinct gene expression profiles associated with each cluster.

C.3 DISCOVERMODE DATASET I

Unexpected antigen presentation in non-immune cells

****1. DATASET EVIDENCE**** Cluster 10: - CCL5 (0.9835575485799701 pct_in, 0.18777651034615098 pct_out)- CD3D (0.8542600896860987 pct_in, 0.1579906826007697 pct_out) - S100A4 (0.851270553064275 pct_in, 0.4219587859663337 pct_out) - CD52 (0.8430493273542601 pct_in, 0.19570802639574428 pct_out) - IL32(0.8266068759342302 pct_in, 0.2018059123908617 pct_out) - CD2 (0.8262331838565022 pct_in, 0.130880675458141 pct_out) - PTPRC (0.8161434977578476 pct_in, 0.2470496679210686 pct_out) - TRAC(0.7518684603886397 pct_in, 0.1437374071191753 pct_out) - ARHGDIB (0.6812406576980568 pct_in, 0.2766009615897146 pct_out) - CD3G (0.6606875934230194 pct_in, 0.10064710083899235 pct_out) Cluster8: - PTPRC (0.9631685166498486 pct_in, 0.23282710911744042 pct_out) - FYN (0.9492936427850656 pct_in, 0.22585526671350592 pct_out) - CCL5 (0.9427346115035318 pct_in, 0.17844673836675135 pct_out)- AOA (0.9392028254288597 pct_in, 0.15170512889801654 pct_out) - MBNL1(0.9281029263370333 pct_in, 0.5080257255580176 pct_out) - SYTL3 (0.9197780020181635 pct_in, 0.3457169107712263 pct_out) - BTG1(0.8958123107971746 pct_in, 0.5309949737880344 pct_out) - CD96 (0.8852169525731585 pct_in, 0.1383667513376209 pct_out) - PTPNC1 (0.8852169525731585 pct_in, 0.3247365292114792 pct_out) - ARHGAP15 (0.8801715438950555 pct_in, 0.20962006161163055 pct_out) Cluster 16: - SRGN(0.9674908424908425 pct_in, 0.2585396892730261 pct_out) - TYROBP (0.9386446886446886 pct_in, 0.05948353571239196 pct_out) - S100A9 (0.9358974358974359 pct_in, 0.14522509146826448 pct_out) -S100A4 (0.9271978021978022 pct_in, 0.42244021422132666 pct_out) -LYZ (0.9212454212454212 pct_in, 0.2217827032186224 pct_out) - HLA-DRA (0.8988095238095238 pct_in, 0.4820616151439631 pct_out) - AIF1(0.8809523809523809 pct_in, 0.04228219948035421 pct_out) - PLAUR (0.8644688644688645 pct_in, 0.1116177952171377 pct_out) - FCER1G (0.853021978021978 pct_in, 0.04689538151545681 pct_out) - S100A8 (0.8402014652014652 pct_in, 0.045601569542393555 pct_out) Cluster 0: - AC022706.1(0.030966600309666 pct_in, 0.016667470693231705 pct_out) - DNAH12 (0.9856226498562265 pct_in, 0.12560905012301607 pct_out) - CFPAP299 (0.9828946398289464 pct_in, 0.10873655265569974 pct_out) - LRR1Q1 (0.9806090098060901 pct_in, 0.10803704954411694 pct_out) - CHST9 (0.9801666298016662 pct_in, 0.5263278498721597 pct_out) - RP1 (0.9739733097397331 pct_in, 0.10774759998070337 pct_out) - RSPH1(0.971245299712453 pct_in, 0.1325075980510396 pct_out) - CAPS (0.9693283196932831 pct_in, 0.3240026050460707 pct_out) - C20orf85 (0.9492737594927376 pct_in, 0.13947850836991654 pct_out) -CRIP1 (0.9463982894639829 pct_in, 0.5 pct_out) Cluster 11: - PRR4 (0.99657664511221 pct_in, 0.19994672349493872 pct_out) - MIR205HG (0.9912514263978699 pct_in, 0.43817794352690465 pct_out) -KRT19 (0.9859262076835299 pct_in, 0.6221310602024507 pct_out) - SLPI (0.9684290604792697 pct_in, 0.6620245071923282 pct_out) - CD9 (0.9676683149486497 pct_in, 0.631006925945658 pct_out) - S100A2(0.9566375047546596 pct_in, 0.4096004262120405 pct_out) - SNHG29 (0.9566375047546596 pct_in, 0.6041768779968034 pct_out) - HSPB1 (0.9509319132750095 pct_in, 0.5403196590303676 pct_out) -LINC00486 (0.9494104222137695 pct_in, 0.33413958444326053 pct_out) - ADIRF (0.9166983643971092pct_in, 0.43643047416089503 pct_out) ****2. ESTABLISHED BIOLOGY**** - CCL5: involved in chemotaxis and immune response - CD3D: part of the T-cell receptor complex, involved in immune response - S100A4: involved in cell proliferation, differentiation, and immune response - CD52: involved in immune response and cell adhesion - IL32: involved in immune response and inflammation - CD2: involved in immune response and cell adhesion - PTPRC: involved in immune response and cell signaling - TRAC: involved in immune response and T-cell development - ARHGDIB: involved in cell signaling and immune response - CD3G: involved in immune response and T-cell development - PTPRC: involved in immune response and cell signaling - FYN:

involved in immune response and cell signaling - CCL5: involved in chemotaxis and immune response - AOA: involved in immune response and inflammation - MBNL1: involved in RNA processing and immune response - SYTL3: involved in cell signaling and immune response - BTG1: involved in cell signaling and immune response - CD96: involved in immune response and cell adhesion - PTPNC1: involved in cell signaling and immune response - ARHGAP15: involved in cell signaling and immune response - SRGN: involved in immune response and cell adhesion - TYROBP: involved in immune response and cell signaling - S100A9: involved in immune response and cell adhesion - S100A4: involved in cell proliferation, differentiation, and immune response - LYZ: involved in immune response and cell adhesion - HLA-DRA: involved in immune response and antigen presentation - AIF1: involved in immune response and cell adhesion - PLAUR: involved in immune response and cell signaling - FCER1G: involved in immune response and cell signaling - S100A8: involved in immune response and cell adhesion - AC022706.1: no established biology - DNAH12: involved in ciliary function and cell signaling - CFAP299: involved in ciliary function and cell signaling - LRRIQ1: involved in cell signaling and ciliary function - CHST9: involved in cell signaling and ciliary function - RP1: involved in ciliary function and cell signaling - RSPH1: involved in ciliary function and cell signaling - CAPS: involved in ciliary function and cell signaling - C20orf85: no established biology - CRIP1: involved in cell signaling and ciliary function - PRR4: involved in immune response and cell signaling - MIR205HG: no established biology - KRT19: involved in cell signaling and epithelial development - SLPI: involved in immune response and cell signaling - CD9: involved in immune response and cell adhesion - S100A2: involved in cell signaling and immune response - SNHG29: no established biology - HSPB1: involved in cell signaling and immune response - LINC00486: no established biology - ADIRF: involved in cell signaling and immune response **3. CONSISTENCY ANALYSIS** - CCL5, CD3D, S100A4, CD52, IL32, CD2, PTPRC, TRAC, ARHGAP15, and CD3G in Cluster 10 are involved in immune response and cell signaling, which is consistent with established biology. - PTPRC, FYN, CCL5, AOA, MBNL1, SYTL3, BTG1, CD96, PTPNC1, and ARHGAP15 in Cluster 8 are involved in immune response and cell signaling, which is consistent with established biology. - SRGN, TYROBP, S100A9, S100A4, LYZ, HLA-DRA, AIF1, PLAUR, FCER1G, and S100A8 in Cluster 16 are involved in immune response and cell adhesion, which is consistent with established biology. - AC022706.1, C20orf85, and CRIP1 in Cluster 0 have no established biology or are involved in ciliary function and cell signaling, which is unexpected. - PRR4, MIR205HG, KRT19, SLPI, CD9, S100A2, SNHG29, HSPB1, LINC00486, and ADIRF in Cluster 11 are involved in immune response and cell signaling, which is consistent with established biology. **4. CANDIDATE NOVEL HYPOTHESES** - It is possible that AC022706.1, C20orf85, and CRIP1 in Cluster 0 are involved in novel mechanisms of immune response or cell signaling, given their enrichment in this cluster. - The presence of PRR4, MIR205HG, KRT19, SLPI, CD9, S100A2, SNHG29, HSPB1, LINC00486, and ADIRF in Cluster 11 suggests that these genes may be involved in novel mechanisms of immune response or cell signaling in epithelial cells. - The unexpected presence of DNAH12, CFAP299, LRRIQ1, CHST9, RP1, RSPH1, and CAPS in Cluster 0 suggests that these genes may be involved in novel mechanisms of ciliary function or cell signaling in non-immune cells.

Discover: Hypoxic signaling in airway basal cells

DATASET EVIDENCE Cluster 12: - FAM155A (logFC: 9.814146995544434, pct_in: 0.05746677091477717, pct_out: 0.024339604561280224) - IGHG1 (logFC: 4.5883684158325195, pct_in: 0.07388584831899922, pct_out: 0.03656264307236933) - TFAA1 (logFC: 4.564707279205322, pct_in: 0.06450351837372947, pct_out: 0.02221015534332045) - RGD2 (logFC: 4.20048189163208, pct_in: 0.054730258014073496, pct_out: 0.019984880910552484) - MZB1 (logFC: 3.0464746952056885, pct_in: 0.0527756059421423, pct_out: 0.018590091672788835) - AC139720.1 (logFC: 2.637744665145874, pct_in: 0.04847537138389367, pct_out: 0.016098636087775897) - AC022706.1 (logFC: 2.240598440170288, pct_in: 0.06489444878811572, pct_out: 0.017418894602910958) - CPNE7 (logFC:

1.4247220754623413, pct_in: 0.045738858483189995, pct_out: 0.013958539623726323)
 - FGFBP2 (logFC: 1.1282931566238403, pct_in: 0.04652071931196247, pct_out:
 0.012734106323399453) - LRR1Q1 (logFC: NaN, pct_in: 0.990226739640344, pct_out:
 0.21001692912128278) Cluster 5: - PTPRC (logFC: NaN, pct_in: 0.902258422806368,
 pct_out: 0.22490859382720116) - ARHGAP15 (logFC: NaN, pct_in: 0.8930025916327287,
 pct_out: 0.19827179200017567) - BTG1 (logFC: NaN, pct_in: 0.8796741947426879,
 pct_out: 0.5261921231485447) - MBNL1 (logFC: NaN, pct_in: 0.8574601999259533,
 pct_out: 0.5055831878520373) - FYN (logFC: NaN, pct_in: 0.8335801554979637,
 pct_out: 0.22129626579707282) - SRGN (logFC: NaN, pct_in: 0.8171047760088856,
 pct_out: 0.24241026823457074) - SKAP1 (logFC: NaN, pct_in: 0.7619400222139948,
 pct_out: 0.31543638898953635) - IL32 (logFC: NaN, pct_in: 0.7284339133654202,
 pct_out: 0.18892804989184975) - CXCR4 (logFC: NaN, pct_in: 0.7256571640133284,
 pct_out: 0.1666501970859822) - ARHGDI1B (logFC: NaN, pct_in: 0.7162162162162162,
 pct_out: 0.26241531890598063) Cluster 2: - Z93241.1 (logFC: 6.87161111831665,
 pct_in: 0.0009705204415868009, pct_out: 0.017577859377124982) - KRT19 (logFC:
 NaN, pct_in: 0.976100934125925, pct_out: 0.5999025341130604) - KRT15 (logFC:
 NaN, pct_in: 0.9653038942132719, pct_out: 0.3395552835577315) - MIR205HG (logFC:
 NaN, pct_in: 0.9627562780541065, pct_out: 0.4056507547939616) - SNHG29 (logFC:
 NaN, pct_in: 0.9252699259978163, pct_out: 0.584681989210753) - KRT17 (logFC:
 NaN, pct_in: 0.925148610942618, pct_out: 0.4054807561539508) - PRSS23 (logFC:
 NaN, pct_in: 0.9210238990658741, pct_out: 0.4488644090847273) - CD9 (logFC:
 NaN, pct_in: 0.9179910226859154, pct_out: 0.6142277528446439) - S100A2 (logFC:
 NaN, pct_in: 0.9002790246269562, pct_out: 0.3800602928509905) - HSPB1 (logFC:
 NaN, pct_in: 0.8805046706296251, pct_out: 0.5207738338093295) Cluster 4: - KRT19
 (logFC: NaN, pct_in: 0.9958736993182634, pct_out: 0.6097354380947143) - SLPI
 (logFC: NaN, pct_in: 0.9756010046645138, pct_out: 0.6516583246246082) - AQP3
 (logFC: NaN, pct_in: 0.971654108360244, pct_out: 0.5088388977503988) - HSPB1
 (logFC: NaN, pct_in: 0.9614280588446358, pct_out: 0.5263736868159067) - CD9
 (logFC: NaN, pct_in: 0.9594546106925009, pct_out: 0.620603927176723) - PRSS23
 (logFC: NaN, pct_in: 0.9581987800502332, pct_out: 0.46044772014740665) - SDK1
 (logFC: NaN, pct_in: 0.9483315392895587, pct_out: 0.5324569605632253) - SERPINB3
 (logFC: NaN, pct_in: 0.9472551130247578, pct_out: 0.3182223200044002) - CHST9
 (logFC: NaN, pct_in: 0.9468963042698242, pct_out: 0.5682525713657115) - SNHG29
 (logFC: NaN, pct_in: 0.9293146752780768, pct_out: 0.5944337495187283) Cluster 14:
 - KRT19 (logFC: NaN, pct_in: 0.9848090277777778, pct_out: 0.6234138571807805) -
 CD9 (logFC: NaN, pct_in: 0.96875, pct_out: 0.6321422882930714) - S100A2 (logFC:
 NaN, pct_in: 0.9526909722222222, pct_out: 0.4115848155030528) - HSPB1 (logFC:
 NaN, pct_in: 0.9505208333333334, pct_out: 0.5417467480753916) - SNHG29 (logFC:
 NaN, pct_in: 0.9427083333333334, pct_out: 0.6057340058401911) - MIR205HG (logFC:
 NaN, pct_in: 0.9366319444444444, pct_out: 0.4414228829307141) - AQP3 (logFC:
 NaN, pct_in: 0.9140625, pct_out: 0.526318024953544) - GAS5 (logFC: NaN, pct_in:
 0.9045138888888888, pct_out: 0.5549455800371649) - KRT15 (logFC: NaN, pct_in:
 0.9019097222222222, pct_out: 0.38056809131935226) - ADIRF (logFC: NaN, pct_in:
 0.8819444444444444, pct_out: 0.43893814706663126) ****ESTABLISHED BIOLOGY****
 Epithelial biology: - KRT19: involved in the regulation of epithelial cell differentiation
 and proliferation - KRT15: involved in the regulation of epithelial cell differentiation and
 proliferation - SLPI: involved in the regulation of epithelial cell function and inflammation -
 AQP3: involved in the regulation of water transport across epithelial cells - HSPB1: involved
 in the regulation of epithelial cell stress response Immune biology: - PTPRC: involved in
 the regulation of T cell activation and signaling - ARHGAP15: involved in the regulation
 of immune cell migration and signaling - BTG1: involved in the regulation of immune cell
 function and differentiation - MBNL1: involved in the regulation of immune cell function
 and differentiation - FYN: involved in the regulation of immune cell signaling and function
 - SRGN: involved in the regulation of immune cell function and inflammation - SKAP1:
 involved in the regulation of immune cell signaling and function - IL32: involved in the
 regulation of immune cell function and inflammation - CXCR4: involved in the regulation of
 immune cell migration and signaling - ARHGDI1B: involved in the regulation of immune cell

function and signaling ****CONSISTENCY ANALYSIS**** MATCHES established biology:
 - Cluster 12: FAM155A, IGHG1, TAF1, RGP2, MZB1, AC139720.1, AC022706.1, CPNE7, FGF2 - Cluster 2: KRT19, KRT15, MIR205HG, SNHG29, KRT17, PRSS23, CD9, S100A2, HSPB1 - Cluster 4: KRT19, SLPI, AQP3, HSPB1, CD9, PRSS23, SDK1, SERPINB3, CHST9, SNHG29 - Cluster 14: KRT19, CD9, S100A2, HSPB1, SNHG29, MIR205HG, AQP3, GAS5, KRT15, ADIRF UNEXPECTED or CONTEXT-SHIFTED:
 - Cluster 12: LRR1 (high expression in epithelial cells, but no established function in epithelial biology) - Cluster 5: PTPRC, ARHGAP15, BTG1, MBNL1, FYN, SRGN, SKAP1, IL32, CXCR4, ARHGAP15 (high expression in immune cells, but no established function in hypoxic signaling) - Cluster 2: Z93241.1 (high expression in epithelial cells, but no established function in epithelial biology) - Cluster 4: SDK1, SERPINB3, CHST9 (high expression in epithelial cells, but no established function in epithelial biology) - Cluster 14: GAS5 (high expression in epithelial cells, but no established function in epithelial biology) ****CANDIDATE NOVEL HYPOTHESES**** Probabilistic hypotheses grounded in mismatches between dataset evidence and established biology: - LRR1 may play a role in epithelial cell function or regulation, but its exact function is unknown. - PTPRC, ARHGAP15, BTG1, MBNL1, FYN, SRGN, SKAP1, IL32, CXCR4, ARHGAP15 may be involved in immune cell function or regulation in response to hypoxia, but their exact roles are unknown. - Z93241.1 may be involved in epithelial cell function or regulation, but its exact function is unknown. - SDK1, SERPINB3, CHST9 may be involved in epithelial cell function or regulation, but their exact functions are unknown. - GAS5 may be involved in epithelial cell function or regulation, but its exact function is unknown. Note: These hypotheses are based on the assumption that the genes with high expression in the clusters are involved in the regulation of epithelial or immune cell function, but their exact roles are unknown.

discover: Crosstalk between immune and secretory epithelial cells

****DATASET EVIDENCE**** Cluster 2: - Z93241.1 (logFC: 6.8716111831665, pct_in: 0.0009705204415868009, pct_out: 0.017577859377124982) - KRT19 (pct_in: 0.976100934125925, pct_out: 0.5999025341130604) - KRT15 (pct_in: 0.9653038942132719, pct_out: 0.3395552835577315) - MIR205HG (pct_in: 0.9627562780541065, pct_out: 0.4056507547939616) - SNHG29 (pct_in: 0.9252699259978163, pct_out: 0.584681989210753) - KRT17 (pct_in: 0.925148610942618, pct_out: 0.4054807561539508) - PRSS23 (pct_in: 0.9210238990658741, pct_out: 0.4488644090847273) - CD9 (pct_in: 0.9179910226859154, pct_out: 0.6142277528446439) - S100A2 (pct_in: 0.9002790246269562, pct_out: 0.3800602928509905) - HSPB1 (pct_in: 0.8805046706296251, pct_out: 0.5207738338093295) Cluster 6: - SLPI (pct_in: 0.9923415661497224, pct_out: 0.6519461734022969) - SCGB3A1 (pct_in: 0.9860233582232434, pct_out: 0.5487639168931358) - SCGB1A1 (pct_in: 0.9674516561363201, pct_out: 0.5682585254668187) - KRT19 (pct_in: 0.9663028910587785, pct_out: 0.6129131235206452) - WFDC2 (pct_in: 0.964771204288723, pct_out: 0.5679297799596739) - BPIFB1 (pct_in: 0.9481141106643691, pct_out: 0.4107894275444902) - MSMB (pct_in: 0.9460080413555428, pct_out: 0.4156986937845183) - PIGR (pct_in: 0.9358606165039249, pct_out: 0.2923862540545279) - CHST9 (pct_in: 0.9201608271108558, pct_out: 0.5712391513982642) - LCN2 (pct_in: 0.903312272640245, pct_out: 0.34682651003769616) Cluster 7: - VIM (pct_in: 0.9648740310077519, pct_out: 0.33207003714090805) - SPARCL1 (pct_in: 0.9629360465116279, pct_out: 0.2526339725612067) - CD74 (pct_in: 0.9617248062015504, pct_out: 0.5843033643382313) - LDB2 (pct_in: 0.8563468992248062, pct_out: 0.03175926627757144) - GNG11 (pct_in: 0.846656976744186, pct_out: 0.05219217983562712) - AQP1 (pct_in: 0.8289728682170543, pct_out: 0.022273716581303938) - EMCN (pct_in: 0.815406976744186, pct_out: 0.009204015116241297) - A2M (pct_in: 0.783187984496124, pct_out: 0.05786618444846293) - RAMP2 (pct_in: 0.7742248062015504, pct_out: 0.05786618444846293)

0.06528353780684562) - ACKR1 (pct.in: 0.7655038759689923, pct.out: 0.02112592175504326) Cluster 14: - KRT19 (pct.in: 0.984809027777778, pct.out: 0.6234138571807805) - CD9 (pct.in: 0.96875, pct.out: 0.6321422882930714) - S100A2 (pct.in: 0.9526909722222222, pct.out: 0.4115848155030528) - HSPB1 (pct.in: 0.9505208333333334, pct.out: 0.5417467480753916) - SNHG29 (pct.in: 0.9427083333333334, pct.out: 0.6057340058401911) - MIR205HG (pct.in: 0.9366319444444444, pct.out: 0.4414228829307141) - AQP3 (pct.in: 0.9140625, pct.out: 0.526318024953544) - GAS5 (pct.in: 0.9045138888888888, pct.out: 0.5549455800371649) - KRT15 (pct.in: 0.9019097222222222, pct.out: 0.38056809131935226) - ADIRF (pct.in: 0.8819444444444444, pct.out: 0.43893814706663126) Cluster 11: - PRR4 (pct.in: 0.99657664511221, pct.out: 0.19994672349493872) - MIR205HG (pct.in: 0.9912514263978699, pct.out: 0.43817794352690465) - KRT19 (pct.in: 0.9859262076835299, pct.out: 0.6221310602024507) - SLPI (pct.in: 0.9684290604792697, pct.out: 0.6620245071923282) - CD9 (pct.in: 0.9676683149486497, pct.out: 0.631006925945658) - S100A2 (pct.in: 0.9566375047546596, pct.out: 0.4096004262120405) - SNHG29 (pct.in: 0.9566375047546596, pct.out: 0.6041768779968034) - HSPB1 (pct.in: 0.9509319132750095, pct.out: 0.5403196590303676) - LINC00486 (pct.in: 0.9494104222137695, pct.out: 0.33413958444326053) - ADIRF (pct.in: 0.9166983643971092, pct.out: 0.43643047416089503) ****ESTABLISHED BIOLOGY****
 Epithelial biology: - KRT19, KRT15, KRT17: Keratin family genes involved in maintaining epithelial cell structure and integrity. - MIR205HG: MicroRNA involved in regulating epithelial cell differentiation and proliferation. - SNHG29: Long non-coding RNA involved in regulating epithelial cell growth and differentiation. - S100A2: Calcium-binding protein involved in regulating epithelial cell proliferation and differentiation. Immune biology: - SLPI: Secretory leukocyte peptidase inhibitor involved in regulating immune cell function and inflammation. - SCGB3A1, SCGB1A1: Secretoglobulin family genes involved in regulating immune cell function and inflammation. - BPIFB1: BPI fold-containing family B member 1 involved in regulating immune cell function and inflammation. - MSMB: Microseminoprotein, beta-1 involved in regulating immune cell function and inflammation. Endothelial biology: - VIM: Vimentin involved in regulating endothelial cell structure and function. - SPARCL1: Secreted protein acidic and rich in cysteine-like 1 involved in regulating endothelial cell function and angiogenesis. - CD74: Major histocompatibility complex class II invariant chain involved in regulating endothelial cell function and immune response. ****CONSISTENCY ANALYSIS**** MATCH: - KRT19, KRT15, KRT17 expression in epithelial clusters (2, 6, 14, 11) is consistent with established epithelial biology. - SLPI, SCGB3A1, SCGB1A1, BPIFB1, MSMB expression in immune clusters (6, 11) is consistent with established immune biology. - VIM, SPARCL1, CD74 expression in endothelial cluster (7) is consistent with established endothelial biology. UNEXPECTED/CONTEXT-SHIFTED: - Z93241.1 expression in epithelial cluster (2) is unexpected and requires further investigation. - MIR205HG, SNHG29, S100A2 expression in immune clusters (6, 11) is unexpected and requires further investigation. - AQP3, GAS5 expression in epithelial cluster (14) is unexpected and requires further investigation. ****CANDIDATE NOVEL HYPOTHESES**** Based on the unexpected expression of genes in immune clusters (6, 11), it is possible that: * There is a novel mechanism of epithelial cell regulation in the immune system, involving the expression of epithelial-specific genes in immune cells. * The immune system is influencing epithelial cell function through the expression of immune-specific genes in epithelial cells. Based on the unexpected expression of genes in epithelial cluster (14), it is possible that: * There is a novel mechanism of epithelial cell regulation involving the expression of AQP3 and GAS5 in epithelial cells. * The expression of AQP3 and GAS5 is involved in regulating epithelial cell function in response to immune system signals. Note: These hypotheses are speculative and require further investigation to confirm or refute them.

C.4 PROMPT DATASET 2

Semantic: Immune activation state heterogeneity across immune cell populations

Based on the provided dataset evidence, we can observe that the immune cell populations are heterogeneous across different clusters. Here's a breakdown of the clusters: 1. **Cluster 18**: This cluster has a high semantic similarity of 0.24311864376068115. The cells in this cluster are primarily mast cells (0.788888888888889) and myeloid cells (1.0). The gene evidence suggests that genes like ARHGAP15, GATA2, and TIMP3 are highly expressed in these cells. 2. **Cluster 3**: This cluster has a high semantic similarity of 0.23854397237300873. The cells in this cluster are primarily macrophages (0.9460269865067467) and myeloid cells (1.0). The gene evidence suggests that genes like HSPA1B, HSPA1A, and SLC9A9 are highly expressed in these cells. 3. **Cluster 16**: This cluster has a high semantic similarity of 0.23643308877944946. The cells in this cluster are primarily macrophages (0.9645390070921985) and myeloid cells (1.0). The gene evidence suggests that genes like GPNMB, NEAT1, and KLF6 are highly expressed in these cells. 4. **Cluster 1**: This cluster has a high semantic similarity of 0.22896818816661835. The cells in this cluster are primarily macrophages (0.7407704654895666) and myeloid cells (1.0). The gene evidence suggests that genes like LGMN, C1QA, and DAB2 are highly expressed in these cells. 5. **Cluster 11**: This cluster has a high semantic similarity of 0.22573009133338928. The cells in this cluster are primarily microglia (0.7148148148148148) and myeloid cells (1.0). The gene evidence suggests that genes like PLXDC2, S100A11, and DOCK4 are highly expressed in these cells. In summary, the immune cell populations are heterogeneous across different clusters, with each cluster having distinct cell types and gene expression profiles. The genes mentioned above are highly expressed in their respective clusters, but their specific roles and functions are not explicitly stated in the provided dataset evidence.

Hybrid: Immune activation state heterogeneity across immune cell populations

Based on the provided dataset evidence, we can observe that there are several clusters of immune cells with varying levels of immune activation state heterogeneity. Here's a brief analysis of each cluster: **Cluster 18**: This cluster consists of 90 cells, primarily classified as Myeloid cells with a subclass of Mast cells (78.89) and hybrid similarity (0.848372042) with other clusters. However, the semantic similarity is relatively low (0.243118643). The cluster does not have any GO or Reactome terms associated with it. **Cluster 3**: This cluster comprises 667 cells, primarily classified as Myeloid cells with a subclass of Macrophages Monocyte-like (94.602699). The cluster shows high expression similarity (0.999672770) and hybrid similarity (0.847447038) with other clusters. However, the semantic similarity is relatively low (0.238543972). The cluster does not have any GO or Reactome terms associated with it. **Cluster 16**: This cluster consists of 141 cells, primarily classified as Myeloid cells with a subclass of Macrophages-iron metabolism (96.453900). The cluster shows high expression similarity (0.999909520) and hybrid similarity (0.847214282) with other clusters. However, the semantic similarity is relatively low (0.236433089). The cluster does not have any GO or Reactome terms associated with it. **Cluster 1**: This cluster comprises 1246 cells, primarily classified as Myeloid cells with a subclass of Macrophages (74.077046). The cluster shows high expression similarity (0.999656379) and hybrid similarity (0.845518768) with other clusters. However, the semantic similarity is relatively low (0.228968188). The cluster does not have any GO or Reactome terms associated with it. **Cluster 11**: This cluster consists of 270 cells, primarily classified as Myeloid cells with a subclass of Microglia (71.481481). The cluster shows high expression similarity (0.999705434) and hybrid similarity (0.844910383) with other clusters. However, the semantic similarity is relatively low (0.225730091). The cluster does not have any GO or Reactome terms associated with it. In summary, these clusters represent different subpopulations of immune cells with varying levels of immune activation state heterogeneity. While the clusters show high expression similarity, the semantic similarity is relatively low, indicating that the clusters may have distinct biological characteristics. However, without further analysis and validation, it is difficult to draw more specific conclusions about the biological relevance of these clusters. The genes mentioned in the dataset evidence are: * ARHGAP15 * GATA2 * TIMP3 * S100A6 * GIHCG * SLC24A3 * MEIS2 * ITGA4 * TAGLN2 * SOX4 * HSPA1B * HSPA1A * SLC9A9 * MRC1 * HSP90AA1 * RPS4Y1 * DNAJB1 * STAB1 * HSPB1 * JUN * GPNMB * NEAT1

* SLC2A1 * KLF6 * FOS * DUSP1 * ATF3 * SH3BP5 * ARL4C * SAMD4A * LGMN * C1QA * DAB2 * FCGRT * BLVRB * PLTP * VSIG4 * A2M * C1QB * CD14 * PLXDC2 * SLC8A1 * S100A11 * FRMD4A * DOCK4 * TREM2 * ALOX5AP * CAPG * LIMS1 * C3

These genes are associated with various biological processes, including immune response, cell differentiation, and protein binding. However, without further analysis and validation, it is difficult to draw more specific conclusions about the biological relevance of these genes in the context of immune activation state heterogeneity.

scGPT: Immune activation state heterogeneity across immune cell populations

Based on the provided dataset evidence, I will analyze the immune activation state heterogeneity across immune cell populations. **Cluster Relevance:** 1. **Cluster 16 (Myeloid):** This cluster consists of 141 cells, primarily classified as Myeloid cells with a high expression of genes such as GPNMB, NEAT1, and SLC2A1. The cluster is also enriched with macrophages and microglia, indicating a high level of immune activation. 2. **Cluster 17 (Lymphoid):** This cluster consists of 102 cells, primarily classified as Lymphoid cells with a high expression of genes such as FYN, BTG1, and CD247. The cluster is also enriched with CD56-bright NK cells and ILC precursors, indicating a high level of immune activation. 3. **Cluster 19 (Myeloid):** This cluster consists of 87 cells, primarily classified as Myeloid cells with a high expression of genes such as RUNX1, LYVE1, and FOXO3. The cluster is also enriched with macrophages and monocytes, indicating a high level of immune activation. 4. **Cluster 8 (Myeloid):** This cluster consists of 342 cells, primarily classified as Myeloid cells with a high expression of genes such as RPS26, C1QC, and DAB2. The cluster is also enriched with monocytes, macrophages, and microglia, indicating a high level of immune activation. 5. **Cluster 11 (Myeloid):** This cluster consists of 270 cells, primarily classified as Myeloid cells with a high expression of genes such as PLXDC2, SLC8A1, and S100A11. The cluster is also enriched with microglia, dendritic cells, and macrophages, indicating a high level of immune activation. **Biological Interpretation:** The provided dataset evidence suggests that there is a high level of immune activation state heterogeneity across immune cell populations. The clusters are primarily composed of Myeloid and Lymphoid cells, indicating a strong immune response. The expression of genes such as GPNMB, NEAT1, and SLC2A1 in Cluster 16, and FYN, BTG1, and CD247 in Cluster 17, suggests a high level of immune activation in these clusters. The clusters also show a high level of heterogeneity in terms of cell type and gene expression, indicating a complex immune response. The presence of macrophages, microglia, and dendritic cells in multiple clusters suggests a coordinated immune response. **Conclusion:** In conclusion, the provided dataset evidence suggests that there is a high level of immune activation state heterogeneity across immune cell populations. The clusters are primarily composed of Myeloid and Lymphoid cells, indicating a strong immune response. The expression of specific genes in each cluster suggests a high level of immune activation, and the presence of multiple cell types in each cluster indicates a complex immune response.

Semantic: Transcriptional programs associated with activated versus resting immune cells

Based on the provided dataset evidence, we can observe that the clusters are primarily composed of myeloid cells, with some variation in subclass and development stage ontology term ID. Cluster 3 (semantic similarity: 0.24018561840057373) is characterized by a high proportion of macrophages and monocytes, with a significant presence of genes involved in iron metabolism (e.g., HSPA1B, HSPA1A). This cluster may represent a subset of macrophages involved in iron homeostasis. Cluster 16 (semantic similarity: 0.23690225183963776) is also enriched for macrophages and monocytes, with a notable presence of genes involved in innate immunity (e.g., GPNMB, NEAT1). This cluster may represent a subset of macrophages involved in immune responses. Cluster 18 (semantic similarity: 0.23557183146476746) is characterized by a high proportion of mast cells, with some presence of genes involved in cell adhesion and migration (e.g., ARHGAP15, ITGA4). This cluster may represent a subset of mast cells involved in tissue repair and remodeling. Cluster 19 (semantic similarity:

0.22687971591949463) is enriched for macrophages and monocytes, with a significant presence of genes involved in cell adhesion and migration (e.g., RUNX1, LYVE1). This cluster may represent a subset of macrophages involved in tissue repair and remodeling. Cluster 1 (semantic similarity: 0.22109606862068176) is the largest cluster, composed primarily of macrophages and monocytes, with some presence of genes involved in innate immunity (e.g., LGMN, C1QA). This cluster may represent a diverse population of myeloid cells involved in immune responses. In summary, the clusters represent distinct subsets of myeloid cells, each with unique gene expression profiles and potential biological functions.

Hybrid: Transcriptional programs associated with activated versus resting immune cells

Based on the provided dataset evidence, we can identify several clusters associated with activated versus resting immune cells. Cluster 3 (n_cells = 667) is primarily composed of macrophages and monocytes, with a high expression of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1) and other genes involved in cellular stress response. This cluster is likely associated with activated immune cells. Cluster 16 (n_cells = 141) is also enriched for macrophages, particularly those involved in iron metabolism. This cluster expresses genes such as GPNMB, NEAT1, and KLF6, which are involved in cellular differentiation and immune function. Cluster 18 (n_cells = 90) is primarily composed of mast cells and expresses genes such as ARHGAP15, GATA2, and S100A6, which are involved in cell signaling and immune response. Cluster 19 (n_cells = 87) is enriched for macrophages and expresses genes such as RUNX1, LYVE1, and FOXO3, which are involved in immune cell development and function. Cluster 1 (n_cells = 1246) is the largest cluster and is primarily composed of macrophages. This cluster expresses genes such as LGMN, C1QA, and DAB2, which are involved in immune cell function and cellular stress response. In summary, these clusters are associated with activated immune cells, including macrophages and monocytes, and express genes involved in cellular stress response, immune cell development, and function.

scGPT: Transcriptional programs associated with activated versus resting immune cells

Based on the provided dataset evidence, we can observe three clusters (16, 17, and 19) that are relevant to the query "Transcriptional programs associated with activated versus resting immune cells." Cluster 16 (Myeloid cells) shows a high expression of genes such as GPNMB, NEAT1, and SLC2A1, which are associated with activated immune cells. This cluster has a high hybrid similarity score of 0.9998902082443237, indicating a strong correlation between the semantic and expression similarities. Cluster 17 (Lymphoid cells) is characterized by the expression of genes such as FYN, BTG1, and CD247, which are involved in lymphocyte activation and function. This cluster has a high hybrid similarity score of 0.9998722672462463, indicating a strong correlation between the semantic and expression similarities. Cluster 19 (Myeloid cells) shows a high expression of genes such as RUNX1, LYVE1, and FOXO3, which are associated with activated immune cells. This cluster has a high hybrid similarity score of 0.9997982978820801, indicating a strong correlation between the semantic and expression similarities. In summary, clusters 16, 17, and 19 are relevant to the query and show a high expression of genes associated with activated immune cells. However, it is essential to note that the provided dataset evidence does not allow us to infer causality or make definitive conclusions about the biological processes involved.

Semantic: Shared transcriptional programs between innate and adaptive immune cells

Based on the provided dataset evidence, we can observe that the clusters are primarily composed of myeloid cells, with some variation in subclass and development stage ontology term ID. Cluster 3 (semantic similarity: 0.2597731947898865) is primarily composed of macrophages and monocytes, with a high expression of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1) and other genes involved in cellular stress response. This cluster may represent a general myeloid cell program. Cluster 18 (semantic similarity: 0.25743579864501953) is composed of mast cells and monocytes, with a high expres-

sion of genes involved in cell adhesion (ARHGAP15, ITGA4) and other cellular processes. This cluster may represent a distinct myeloid cell program. Cluster 19 (semantic similarity: 0.24654093384742737) is primarily composed of macrophages and monocytes, with a high expression of genes involved in cell differentiation (RUNX1, LYVE1) and other cellular processes. This cluster may represent a specific myeloid cell program. Cluster 16 (semantic similarity: 0.2452763170003891) is primarily composed of macrophages with iron metabolism, with a high expression of genes involved in cellular metabolism (GPNMB, SLC2A1) and other cellular processes. This cluster may represent a specific myeloid cell program. Cluster 8 (semantic similarity: 0.22802788019180298) is primarily composed of monocytes, with a high expression of genes involved in cellular stress response (RPS26, C1QC) and other cellular processes. This cluster may represent a general myeloid cell program. In summary, the clusters represent distinct myeloid cell programs, with some variation in subclass and development stage ontology term ID. The genes expressed in each cluster are involved in various cellular processes, including stress response, cell adhesion, differentiation, and metabolism.

Hybrid: Shared transcriptional programs between innate and adaptive immune cells

Based on the provided dataset evidence, we can identify clusters that share transcriptional programs between innate and adaptive immune cells. Cluster 3 (n_cells = 667) is primarily composed of macrophages and monocytes, with a high expression similarity (0.9996716976165771) to the query. This cluster is relevant biologically as macrophages are a key component of the innate immune system, while monocytes can differentiate into macrophages or dendritic cells, which are essential for adaptive immune responses. Cluster 8 (n_cells = 342) is also primarily composed of monocytes and macrophages, with a high expression similarity (0.9997067451477051) to the query. This cluster is relevant biologically as monocytes and macrophages play critical roles in both innate and adaptive immune responses. Cluster 16 (n_cells = 141) is primarily composed of macrophages-iron metabolism cells, with a high expression similarity (0.999884843826294) to the query. This cluster is relevant biologically as macrophages-iron metabolism cells are specialized macrophages that play a crucial role in iron homeostasis and immune responses. Cluster 18 (n_cells = 90) is primarily composed of mast cells, with a moderate expression similarity (0.9997135996818542) to the query. This cluster is relevant biologically as mast cells are a type of immune cell that plays a key role in allergic responses and adaptive immunity. Cluster 19 (n_cells = 87) is primarily composed of macrophages and monocytes, with a moderate expression similarity (0.9997967481613159) to the query. This cluster is relevant biologically as macrophages and monocytes play critical roles in both innate and adaptive immune responses. In terms of specific genes, we can see that clusters 3, 8, and 16 have high expression levels of genes such as HSPA1B, HSPA1A, and SLC9A9, which are involved in heat shock protein function and ion transport, respectively. Cluster 18 has high expression levels of genes such as ARHGAP15 and GATA2, which are involved in GTPase activity and transcriptional regulation, respectively. Cluster 19 has high expression levels of genes such as RUNX1 and LYVE1, which are involved in transcriptional regulation and hyaluronan binding, respectively. Overall, these clusters share transcriptional programs between innate and adaptive immune cells, highlighting the complex and dynamic nature of immune responses.

scGPT: Shared transcriptional programs between innate and adaptive immune cells

Based on the provided dataset evidence, we can observe that there are four clusters (16, 17, 19, and 18) with high hybrid similarity (≥ 0.99) between innate and adaptive immune cells. Cluster 16 is primarily composed of macrophages, with a high expression of genes such as GPNMB, NEAT1, and SLC2A1. These cells are involved in innate immunity and play a crucial role in the clearance of pathogens and debris. Cluster 17 is predominantly composed of lymphoid cells, including CD56-bright NK cells and ILC precursors. The expression of genes such as FYN, BTG1, and CD247 suggests a role in adaptive immunity, particularly in the activation and proliferation of lymphocytes. Cluster 19 is primarily composed of myeloid cells, including macrophages and monocytes. The expression of genes such as RUNX1,

LYVE1, and FOXO3 suggests a role in innate immunity, particularly in the regulation of immune responses and tissue repair. Cluster 18 is primarily composed of myeloid cells, including mast cells and monocytes. The expression of genes such as ARHGAP15, GATA2, and TIMP3 suggests a role in innate immunity, particularly in the regulation of immune responses and tissue repair. In summary, these clusters highlight the shared transcriptional programs between innate and adaptive immune cells, with a focus on the regulation of immune responses and tissue repair. However, it is essential to note that these clusters are not mutually exclusive, and individual cells may exhibit characteristics from multiple clusters. The provided dataset evidence suggests that: * Macrophages (Cluster 16) and myeloid cells (Clusters 19 and 18) share transcriptional programs involved in innate immunity. * Lymphoid cells (Cluster 17) share transcriptional programs involved in adaptive immunity. * There is a significant overlap between innate and adaptive immune cells, highlighting the complexity of the immune system. These findings are based solely on the provided dataset evidence and should be interpreted with caution. Further analysis and validation are necessary to confirm these results and explore their biological relevance.

Semantic: Innate-like signaling pathways expressed in adaptive immune populations

Based on the provided dataset evidence, we can observe that cluster 3 is the most relevant to innate-like signaling pathways expressed in adaptive immune populations. This cluster has a semantic similarity of 0.265159010887146 and is characterized by a high proportion of myeloid cells, particularly macrophages and monocytes. The gene evidence for cluster 3 includes genes such as HSPA1B, HSPA1A, SLC9A9, and MRC1, which are involved in heat shock protein activity and cellular stress response. These genes are expressed at high levels in macrophages and monocytes, suggesting that they play a role in the innate-like signaling pathways of these cells. Cluster 18 is also relevant, with a semantic similarity of 0.24977195262908936. This cluster is characterized by a high proportion of mast cells and myeloid cells, and includes genes such as ARHGAP15, GATA2, and TIMP3, which are involved in signaling pathways and cellular differentiation. Cluster 2 has a semantic similarity of 0.24531462788581848 and is characterized by a high proportion of macrophages and myeloid cells. The gene evidence for this cluster includes genes such as RPS4Y1, HSPA1A, and HSPA1B, which are involved in ribosomal protein activity and heat shock protein activity. In summary, the most relevant clusters to innate-like signaling pathways expressed in adaptive immune populations are cluster 3, cluster 18, and cluster 2. These clusters are characterized by a high proportion of myeloid cells, particularly macrophages and monocytes, and include genes involved in signaling pathways, cellular stress response, and cellular differentiation.

Hybrid: Innate-like signaling pathways expressed in adaptive immune populations

Based on the provided dataset evidence, I will analyze the clusters related to innate-like signaling pathways expressed in adaptive immune populations. Cluster 3 (hybrid_similarity: 0.85277259349823) is primarily composed of macrophages and monocytes, with a high expression of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1, HSPB1) and other genes involved in cellular stress response. This cluster may represent a population of myeloid cells that exhibit innate-like characteristics, such as enhanced stress response and antigen presentation capabilities. Cluster 18 (hybrid_similarity: 0.8497061133384705) is composed of mast cells and a smaller fraction of monocytes. This cluster expresses genes involved in cell adhesion (ARHGAP15, ITGA4) and signaling pathways (GATA2, MEIS2). The presence of mast cells in this cluster suggests that this population may be involved in innate-like immune responses, such as allergic reactions or tissue repair. Cluster 2 (hybrid_similarity: 0.8481488823890686) is primarily composed of macrophages, with a high expression of heat shock proteins (HSPA1A, HSPA1B, HSPB1) and other genes involved in cellular stress response. This cluster may represent a population of myeloid cells that exhibit innate-like characteristics, similar to Cluster 3. Cluster 8 (hybrid_similarity: 0.8439385294914246) is composed of monocytes and a smaller fraction of macrophages. This cluster expresses

genes involved in cellular stress response (RPS26, C1QC, RPL39) and other functions. The presence of monocytes in this cluster suggests that this population may be involved in innate-like immune responses, such as inflammation or tissue repair. Cluster 15 (hybrid_similarity: 0.8439268469810486) is composed of megakaryocytes and platelets. This cluster expresses genes involved in cell adhesion (CNST, LIMS1) and signaling pathways (RGS18, PBX1). The presence of megakaryocytes and platelets in this cluster suggests that this population may be involved in innate-like immune responses, such as coagulation or vascular repair. In summary, these clusters represent populations of myeloid cells that exhibit innate-like characteristics, such as enhanced stress response, antigen presentation capabilities, and involvement in innate-like immune responses. The specific genes and cell types present in each cluster provide insights into the biological relevance of these populations.

scGPT: Innate-like signaling pathways expressed in adaptive immune populations

Based on the provided dataset evidence, we can analyze the clusters related to innate-like signaling pathways expressed in adaptive immune populations. Cluster 17 (Lymphoid) shows a high expression of genes involved in immune responses, such as FYN, BTG1, and CD247. These genes are typically associated with lymphoid cells, suggesting that this cluster may represent a subset of adaptive immune cells with innate-like characteristics. Cluster 19 (Myeloid) exhibits a high expression of genes related to myeloid cells, including RUNX1, LYVE1, and AGR2. While these genes are not typically associated with adaptive immune responses, their expression in this cluster may indicate a myeloid cell subset with innate-like signaling pathways. Cluster 8 (Myeloid) shows a high expression of genes involved in immune responses, such as RPS26, C1QC, and DAB2. These genes are typically associated with myeloid cells, suggesting that this cluster may represent a subset of myeloid cells with innate-like characteristics. Cluster 18 (Myeloid) exhibits a high expression of genes related to mast cells, including ARHGAP15, GATA2, and TIMP3. While these genes are not typically associated with adaptive immune responses, their expression in this cluster may indicate a mast cell subset with innate-like signaling pathways. In summary, clusters 17, 19, 8, and 18 show a high expression of genes involved in immune responses, suggesting that they may represent subsets of adaptive immune cells or myeloid cells with innate-like characteristics. However, further analysis is needed to confirm the exact nature of these clusters and their relationship to innate-like signaling pathways.

Semantic: Transcriptional signatures of immune exhaustion or dysfunction

Based on the provided dataset evidence, I will analyze the transcriptional signatures of immune exhaustion or dysfunction. The dataset contains 6 clusters with varying semantic similarities to the query "Transcriptional signatures of immune exhaustion or dysfunction". The clusters are: 1. Cluster 20: This cluster has a semantic similarity of 0.1445765644311905 and consists of 65 cells. The cluster is primarily composed of lymphoid cells, with a specific subclass of ILC precursors. The gene evidence suggests that this cluster is enriched for genes involved in immune cell development and function, such as CD52, TRBC1, and IL7R. 2. Cluster 1: This cluster has a semantic similarity of 0.14037597179412842 and consists of 1246 cells. The cluster is primarily composed of myeloid cells, with a specific subclass of macrophages. The gene evidence suggests that this cluster is enriched for genes involved in immune cell function and inflammation, such as LGMN, C1QA, and DAB2. 3. Cluster 19: This cluster has a semantic similarity of 0.13394546508789062 and consists of 87 cells. The cluster is primarily composed of myeloid cells, with a specific subclass of macrophages. The gene evidence suggests that this cluster is enriched for genes involved in immune cell function and inflammation, such as RUNX1, LYVE1, and A2M. 4. Cluster 7: This cluster has a semantic similarity of 0.11493303626775742 and consists of 367 cells. The cluster is primarily composed of myeloid cells, with a specific subclass of dendritic cells. The gene evidence suggests that this cluster is enriched for genes involved in immune cell function and antigen presentation, such as CD74, VIM, and HLA-DRA. 5. Cluster 0: This cluster has a semantic similarity of 0.11174975335597992 and consists of 1423 cells. The cluster

is primarily composed of myeloid cells, with a specific subclass of macrophages. The gene evidence suggests that this cluster is enriched for genes involved in immune cell function and inflammation, such as FOS, RNASE1, and FOLR2. Based on the gene evidence, it appears that clusters 1, 19, and 0 are enriched for genes involved in immune cell function and inflammation, which may be indicative of immune exhaustion or dysfunction. However, it is essential to note that the semantic similarity of these clusters to the query is relatively low, and further analysis is required to confirm the relevance of these clusters to immune exhaustion or dysfunction. In conclusion, the provided dataset evidence suggests that clusters 1, 19, and 0 may be related to immune exhaustion or dysfunction, but further analysis is required to confirm this association.

Hybrid: Transcriptional signatures of immune exhaustion or dysfunction

Based on the provided dataset evidence, the clusters can be analyzed as follows: Cluster 20: - This cluster has a high hybrid similarity of 0.8286760449409485. - The cells in this cluster are primarily lymphoid cells, specifically ILC precursors. - The development stage ontology term ID suggests that these cells are in an early developmental stage. - The donor ID and dissection fields indicate that these cells are from various donors and dissections, but the majority are from the meninges. - Genes such as CD52, TRBC1, and ID2 are highly expressed in these cells. Cluster 1: - This cluster has a high hybrid similarity of 0.8278312087059021. - The cells in this cluster are primarily myeloid cells, specifically macrophages and monocytes. - The development stage ontology term ID suggests that these cells are in a mature developmental stage. - The donor ID and dissection fields indicate that these cells are from various donors and dissections, but the majority are from the meninges. - Genes such as LGMN, C1QA, and DAB2 are highly expressed in these cells. Cluster 19: - This cluster has a moderate hybrid similarity of 0.8265514969825745. - The cells in this cluster are primarily myeloid cells, specifically macrophages and monocytes. - The development stage ontology term ID suggests that these cells are in an early developmental stage. - The donor ID and dissection fields indicate that these cells are from two donors and dissections, but the majority are from the meninges. - Genes such as RUNX1, LYVE1, and RPS4Y1 are highly expressed in these cells. Cluster 0: - This cluster has a moderate hybrid similarity of 0.8217591643333435. - The cells in this cluster are primarily myeloid cells, specifically macrophages and monocytes. - The development stage ontology term ID suggests that these cells are in a mature developmental stage. - The donor ID and dissection fields indicate that these cells are from various donors and dissections, but the majority are from the meninges. - Genes such as FOS, RNASE1, and FOLR2 are highly expressed in these cells. Cluster 3: - This cluster has a moderate hybrid similarity of 0.8217124342918396. - The cells in this cluster are primarily myeloid cells, specifically macrophages and monocytes. - The development stage ontology term ID suggests that these cells are in a mature developmental stage. - The donor ID and dissection fields indicate that these cells are from one donor and dissection, but the majority are from the meninges. - Genes such as HSPA1B, HSPA1A, and SLC9A9 are highly expressed in these cells. In summary, the clusters can be broadly categorized into two groups: lymphoid cells (Cluster 20) and myeloid cells (Clusters 1, 19, 0, and 3). The myeloid cells can be further divided into macrophages and monocytes, with varying developmental stages and expression profiles.

scGPT: Transcriptional signatures of immune exhaustion or dysfunction

Based on the provided dataset evidence, we can analyze the transcriptional signatures of immune exhaustion or dysfunction in the following clusters: Cluster 16: This cluster has a high hybrid similarity of 0.9999071359634399 and is primarily composed of myeloid cells, specifically macrophages and microglia. The genes GPNMB, NEAT1, and KLF6 are highly expressed in this cluster. These genes are involved in various cellular processes, including cell adhesion, migration, and differentiation. However, their specific roles in immune exhaustion or dysfunction are not explicitly stated in the provided evidence. Cluster 17: This cluster has a high hybrid similarity of 0.9998222589492798 and is primarily composed of

lymphoid cells, specifically CD56-bright NK cells and ILC precursors. The genes FYN, BTG1, and CD247 are highly expressed in this cluster. These genes are involved in various immune-related processes, including T-cell activation and differentiation. However, their specific roles in immune exhaustion or dysfunction are not explicitly stated in the provided evidence. Cluster 11: This cluster has a high hybrid similarity of 0.9997247457504272 and is primarily composed of myeloid cells, specifically microglia and dendritic cells. The genes PLXDC2, S100A11, and TREM2 are highly expressed in this cluster. These genes are involved in various immune-related processes, including antigen presentation and recognition. However, their specific roles in immune exhaustion or dysfunction are not explicitly stated in the provided evidence. Cluster 19: This cluster has a high hybrid similarity of 0.999703049659729 and is primarily composed of myeloid cells, specifically macrophages and monocytes. The genes RUNX1, LYVE1, and FOXO3 are highly expressed in this cluster. These genes are involved in various cellular processes, including cell differentiation and survival. However, their specific roles in immune exhaustion or dysfunction are not explicitly stated in the provided evidence. Cluster 20: This cluster has a high hybrid similarity of 0.9997009634971619 and is primarily composed of lymphoid cells, specifically ILC precursors. The genes CD52, TRBC1, and ID2 are highly expressed in this cluster. These genes are involved in various immune-related processes, including T-cell activation and differentiation. However, their specific roles in immune exhaustion or dysfunction are not explicitly stated in the provided evidence. In summary, while the provided dataset evidence suggests that these clusters are involved in immune-related processes, the specific roles of the genes mentioned in immune exhaustion or dysfunction are not explicitly stated. Further analysis and investigation are required to determine the exact mechanisms and relationships between these clusters and immune exhaustion or dysfunction.

Semantic: Regulatory immune programs beyond canonical Treg annotations

Based on the provided dataset evidence, I will analyze the clusters and provide a biological explanation for their relevance. **Cluster 18** This cluster has a semantic similarity of 0.20669732987880707 and consists of 90 cells. The metadata indicates that these cells are primarily Myeloid cells, with a high proportion of Mast cells (0.7888888888888889). The gene evidence shows a high expression of genes such as ARHGAP15, GATA2, and S100A6, which are involved in cell signaling, transcriptional regulation, and immune response. **Cluster 11** This cluster has a semantic similarity of 0.1965542882680893 and consists of 270 cells. The metadata indicates that these cells are primarily Myeloid cells, with a high proportion of Microglia (0.7148148148148148). The gene evidence shows a high expression of genes such as PLXDC2, SLC8A1, and S100A11, which are involved in cell signaling, ion transport, and immune response. **Cluster 10** This cluster has a semantic similarity of 0.1940487176179886 and consists of 291 cells. The metadata indicates that these cells are primarily Myeloid cells, with a high proportion of Macrophages Monocyte-like cells (0.9312714776632303). The gene evidence shows a high expression of genes such as SRGN, RNASE1, and LYVE1, which are involved in cell signaling, proteolysis, and immune response. **Cluster 15** This cluster has a semantic similarity of 0.19282454252243042 and consists of 146 cells. The metadata indicates that these cells are primarily Myeloid cells, with a high proportion of Megakaryocytes (0.684931506849315). The gene evidence shows a high expression of genes such as CNST, LTBP1, and LIMS1, which are involved in cell signaling, cell adhesion, and immune response. **Cluster 16** This cluster has a semantic similarity of 0.1912381500005722 and consists of 141 cells. The metadata indicates that these cells are primarily Myeloid cells, with a high proportion of Macrophages-iron metabolism cells (0.9645390070921985). The gene evidence shows a high expression of genes such as GPNMB, NEAT1, and SLC2A1, which are involved in cell signaling, transcriptional regulation, and immune response. In summary, these clusters are primarily composed of Myeloid cells, with varying proportions of different subtypes. The gene evidence suggests that these cells are involved in cell signaling, transcriptional regulation, immune response, and proteolysis. However, it is essential to note that these findings are based solely on the provided dataset evidence and should be further validated through additional experiments and analyses.

Hybrid: Regulatory immune programs beyond canonical Treg annotations

Based on the provided dataset evidence, I will analyze the clusters and provide a biological explanation for each. **Cluster 18** This cluster has a high hybrid similarity score of 0.8411178588867188. The metadata indicates that the cells in this cluster are primarily Myeloid cells (ClassAnn: Myeloid = 1.0). The subclass annotation shows a mix of Mast cells, MEMP-MEP, and Monocytes. The development stage ontology term ID suggests that these cells are at various stages of development. The gene evidence shows a high percentage of cells expressing genes such as ARHGAP15, GATA2, TIMP3, and S100A6, which are involved in various cellular processes. However, without further information on the specific functions of these genes in Myeloid cells, it is difficult to provide a more detailed explanation. **Cluster 11** This cluster has a high hybrid similarity score of 0.8390102982521057. The metadata indicates that the cells in this cluster are primarily Myeloid cells (ClassAnn: Myeloid = 1.0). The subclass annotation shows a mix of Microglia, Dendritic cells, and Macrophages. The development stage ontology term ID suggests that these cells are at various stages of development. The gene evidence shows a high percentage of cells expressing genes such as PLXDC2, SLC8A1, S100A11, and FRMD4A, which are involved in various cellular processes. However, without further information on the specific functions of these genes in Myeloid cells, it is difficult to provide a more detailed explanation. **Cluster 15** This cluster has a high hybrid similarity score of 0.8383318781852722. The metadata indicates that the cells in this cluster are primarily Myeloid cells (ClassAnn: Myeloid = 1.0). The subclass annotation shows a mix of Megakaryocytes, Platelets, and MEMP-MEP. The development stage ontology term ID suggests that these cells are at various stages of development. The gene evidence shows a high percentage of cells expressing genes such as CNST, LTBP1, LIMS1, and RGS18, which are involved in various cellular processes. However, without further information on the specific functions of these genes in Myeloid cells, it is difficult to provide a more detailed explanation. **Cluster 16** This cluster has a high hybrid similarity score of 0.8381441235542297. The metadata indicates that the cells in this cluster are primarily Myeloid cells (ClassAnn: Myeloid = 1.0). The subclass annotation shows a mix of Macrophages, Microglia, and Macrophages Monocyte-like. The development stage ontology term ID suggests that these cells are at various stages of development. The gene evidence shows a high percentage of cells expressing genes such as GPNMB, NEAT1, SLC2A1, and KLF6, which are involved in various cellular processes. However, without further information on the specific functions of these genes in Myeloid cells, it is difficult to provide a more detailed explanation. **Cluster 19** This cluster has a high hybrid similarity score of 0.8377711772918701. The metadata indicates that the cells in this cluster are primarily Myeloid cells (ClassAnn: Myeloid = 1.0). The subclass annotation shows a mix of Macrophages Monocyte-like, Macrophages, and Monocytes. The development stage ontology term ID suggests that these cells are at various stages of development. The gene evidence shows a high percentage of cells expressing genes such as RUNX1, LYVE1, RPS4Y1, and FOXO3, which are involved in various cellular processes. However, without further information on the specific functions of these genes in Myeloid cells, it is difficult to provide a more detailed explanation. In summary, these clusters appear to represent different subpopulations of Myeloid cells, each with unique characteristics and gene expression profiles. However, further analysis is needed to understand the specific biological functions of these cells and the genes involved.

scGPT: Regulatory immune programs beyond canonical Treg annotations

Based on the provided dataset evidence, we can observe that there are four clusters (16, 17, 18, and 19) that are primarily annotated as Myeloid cells. These clusters have high hybrid similarity (≥ 0.99) and are characterized by distinct gene expression profiles. Cluster 16 is primarily composed of Macrophages-iron metabolism and Macrophages Monocyte-like cells. The gene evidence for this cluster includes GPNMB, NEAT1, and SLC2A1, which are involved in iron metabolism and cellular stress responses. Cluster 17 is primarily composed of CD56-bright NK cells and ILC precursors. The gene evidence for this cluster includes FYN, BTG1, and CD247, which are involved in immune cell signaling and regulation. Clus-

ter 18 is primarily composed of Mast cells, MEMP-MEP, and Monocytes. The gene evidence for this cluster includes ARHGAP15, GATA2, and TIMP3, which are involved in cell signaling, transcriptional regulation, and cellular processes. Cluster 19 is primarily composed of Megakaryocytes, Platelets, and MEMP-MEP. The gene evidence for this cluster includes CNST, LTBP1, and LIMS1, which are involved in cellular processes and signaling pathways. In summary, these clusters represent distinct regulatory immune programs beyond canonical Treg annotations, with each cluster exhibiting unique gene expression profiles and cellular characteristics.

Semantic: Interferon-stimulated gene programs across immune cell types

Based on the provided dataset evidence, I will analyze the interferon-stimulated gene programs across immune cell types. ****Cluster Relevance:**** Cluster 18 (semantic similarity: 0.2659740447998047) is associated with Myeloid cells, specifically Mast cells (0.7888888888888889), MEMP-MEP (0.2), and Monocytes (0.0111111111111112). This cluster shows a high expression of genes such as ARHGAP15, GATA2, TIMP3, and S100A6, which are involved in various cellular processes including cell migration, differentiation, and immune response. Cluster 3 (semantic similarity: 0.2643201947212219) is also associated with Myeloid cells, specifically Macrophages Monocyte-like (0.9460269865067467), Macrophages-iron metabolism (0.034482758620689655), and Monocytes (0.0014992503748125937). This cluster shows a high expression of genes such as HSPA1B, HSPA1A, SLC9A9, and MRC1, which are involved in heat shock response, protein folding, and immune response. Cluster 16 (semantic similarity: 0.24742969870567322) is associated with Myeloid cells, specifically Macrophages-iron metabolism (0.9645390070921985), Macrophages Monocyte-like (0.014184397163120567), and Microglia (0.0070921985815602835). This cluster shows a high expression of genes such as GPNMB, NEAT1, SLC2A1, and KLF6, which are involved in cellular processes including cell migration, differentiation, and immune response. Cluster 2 (semantic similarity: 0.23518583178520203) is associated with Myeloid cells, specifically Macrophages-iron metabolism (0.8507306889352818), Macrophages Monocyte-like (0.10751565762004175), and Macrophages (0.03966597077244259). This cluster shows a high expression of genes such as RPS4Y1, HSPA1A, HSPA1B, and SELENOP, which are involved in cellular processes including protein synthesis, heat shock response, and immune response. Cluster 13 (semantic similarity: 0.23401576280593872) is associated with Myeloid cells, specifically ChP Microglia (0.9939393939393939) and Dendritic cells (0.0060606060606061). This cluster shows a high expression of genes such as TTR, CD74, PLXDC2, and RPL41, which are involved in cellular processes including cell migration, differentiation, and immune response. ****Interferon-Stimulated Gene Programs:**** The provided dataset evidence shows that interferon-stimulated gene programs are present across various immune cell types, including Myeloid cells. The expression of genes such as HSPA1B, HSPA1A, SLC9A9, and MRC1 in Cluster 3, and RPS4Y1, HSPA1A, HSPA1B, and SELENOP in Cluster 2, suggests that these cells are responding to interferon stimulation. Similarly, the expression of genes such as GPNMB, NEAT1, SLC2A1, and KLF6 in Cluster 16, and TTR, CD74, PLXDC2, and RPL41 in Cluster 13, also suggests that these cells are responding to interferon stimulation. However, it is essential to note that the expression of these genes does not necessarily imply that they are directly involved in interferon-stimulated gene programs. Further analysis and experimental validation are required to determine the specific roles of these genes in interferon-stimulated gene programs across immune cell types.

Hybrid: Interferon-stimulated gene programs across immune cell types

Interferon-stimulated gene programs across immune cell typesBased on the provided dataset evidence, we can observe that there are several clusters that exhibit interferon-stimulated gene programs across immune cell types. Here's a brief analysis of each relevant cluster: ****Cluster 18****: This cluster has a high hybrid similarity score of 0.8529582023620605 and is primarily composed

of mast cells (88.89%). The gene evidence suggests that this cluster is enriched for genes involved in interferon-stimulated responses, such as ARHGAP15, GATA2, and S100A6. These genes are often upregulated in response to interferon signaling, which is a key component of the immune response. **Cluster 3**: This cluster has a high hybrid similarity score of 0.8526052236557007 and is primarily composed of macrophages (94.62%). The gene evidence suggests that this cluster is enriched for genes involved in heat shock protein responses, such as HSPA1B, HSPA1A, and HSP90AA1. These genes are often upregulated in response to stress and heat shock, which can be triggered by interferon signaling. **Cluster 16**: This cluster has a high hybrid similarity score of 0.8494004011154175 and is primarily composed of macrophages (96.45%). The gene evidence suggests that this cluster is enriched for genes involved in interferon-stimulated responses, such as GPNMB, NEAT1, and KLF6. These genes are often upregulated in response to interferon signaling, which is a key component of the immune response. **Cluster 2**: This cluster has a high hybrid similarity score of 0.8460900783538818 and is primarily composed of macrophages (85.07%). The gene evidence suggests that this cluster is enriched for genes involved in heat shock protein responses, such as HSPA1A, HSPA1B, and HSPB1. These genes are often upregulated in response to stress and heat shock, which can be triggered by interferon signaling. **Cluster 13**: This cluster has a high hybrid similarity score of 0.8460356593132019 and is primarily composed of microglia (99.39%). The gene evidence suggests that this cluster is enriched for genes involved in interferon-stimulated responses, such as TTR, CD74, and PLXDC2. These genes are often upregulated in response to interferon signaling, which is a key component of the immune response. In summary, these clusters exhibit interferon-stimulated gene programs across immune cell types, with enrichment for genes involved in heat shock protein responses and interferon-stimulated responses. However, it's essential to note that these findings are based solely on the provided dataset evidence and should be interpreted with caution.

scGPT: Interferon-stimulated gene programs across immune cell types

Based on the provided dataset evidence, we can observe three distinct clusters of immune cells that exhibit interferon-stimulated gene programs. **Cluster 16 (Myeloid cells)** This cluster consists of 141 cells, primarily classified as Myeloid cells. The genes GPNMB, NEAT1, SLC2A1, KLF6, FOS, DUSP1, ATF3, SH3BP5, ARL4C, and SAMD4A are upregulated in these cells. These genes are associated with various cellular processes, including cell adhesion, metabolism, and stress response. The presence of these genes suggests that Cluster 16 is involved in the interferon-stimulated gene program, which is a key response to viral infections. **Cluster 17 (Lymphoid cells)** This cluster consists of 102 cells, primarily classified as Lymphoid cells. The genes FYN, BTG1, CD247, GZMA, NKG7, PRKCH, CD69, CD44, GZMM, and KLRB1 are upregulated in these cells. These genes are associated with T-cell activation, proliferation, and cytotoxicity. The presence of these genes suggests that Cluster 17 is involved in the interferon-stimulated gene program, which is a key response to viral infections. **Cluster 19 (Myeloid cells)** This cluster consists of 87 cells, primarily classified as Myeloid cells. The genes RUNX1, LYVE1, RPS4Y1, FOXO3, CEMIP2, AGR2, PLD4, A2M, ZEB2, and NEAT1 are upregulated in these cells. These genes are associated with various cellular processes, including cell adhesion, metabolism, and stress response. The presence of these genes suggests that Cluster 19 is involved in the interferon-stimulated gene program, which is a key response to viral infections. **Cluster 18 (Myeloid cells)** This cluster consists of 90 cells, primarily classified as Myeloid cells. The genes ARHGAP15, GATA2, TIMP3, S100A6, GIHCG, SLC24A3, MEIS2, ITGA4, TAGLN2, and SOX4 are upregulated in these cells. These genes are associated with various cellular processes, including cell adhesion, metabolism, and stress response. The presence of these genes suggests that Cluster 18 is involved in the interferon-stimulated gene program, which is a key response to viral infections. **Cluster 8 (Myeloid cells)** This cluster consists of 342 cells, primarily classified as Myeloid cells. The genes RPS26, C1QC, RPL39, DAB2, CST3, AGR2, RNASE1, RPL41, FTL, and BLVRB are upregulated in these cells. These genes are associated with various cellular processes, including cell adhesion, metabolism, and stress response. The presence of these genes suggests that Cluster 8 is in-

involved in the interferon-stimulated gene program, which is a key response to viral infections. In summary, the provided dataset evidence suggests that there are multiple clusters of immune cells that exhibit interferon-stimulated gene programs. These clusters are primarily composed of Myeloid cells, with some Lymphoid cells present in Cluster 17. The upregulated genes in each cluster are associated with various cellular processes, including cell adhesion, metabolism, and stress response.

Sematinc: Inflammatory signaling heterogeneity within immune populations

Based on the provided dataset evidence, I will analyze the inflammatory signaling heterogeneity within immune populations. **Cluster Relevance:** Cluster 3 (n_cells = 667) is predominantly composed of macrophages and monocyte-like cells, with a high semantic similarity of 0.2945103347301483. This cluster is likely involved in inflammatory signaling, as evidenced by the presence of genes such as HSPA1B, HSPA1A, and SLC9A9, which are involved in heat shock protein and ion transport functions. Cluster 18 (n_cells = 90) is primarily composed of mast cells, with a moderate semantic similarity of 0.262997031211853. This cluster may be involved in inflammatory responses, as indicated by the presence of genes such as ARHGAP15, GATA2, and TIMP3, which are involved in signaling pathways and matrix metalloproteinase regulation. Cluster 16 (n_cells = 141) is mainly composed of macrophages-iron metabolism cells, with a moderate semantic similarity of 0.2568405866622925. This cluster may be involved in inflammatory signaling, as suggested by the presence of genes such as GPNMB, NEAT1, and KLF6, which are involved in cell adhesion and transcriptional regulation. Cluster 2 (n_cells = 958) is predominantly composed of macrophages-iron metabolism cells, with a moderate semantic similarity of 0.23532789945602417. This cluster is likely involved in inflammatory signaling, as evidenced by the presence of genes such as RPS4Y1, HSPA1A, and HSPA1B, which are involved in ribosomal protein and heat shock protein functions. Cluster 8 (n_cells = 342) is primarily composed of monocytes, with a low semantic similarity of 0.2226094901561737. This cluster may be involved in inflammatory responses, as indicated by the presence of genes such as RPS26, C1QC, and DAB2, which are involved in ribosomal protein and complement system regulation. **Biological Interpretation:** The provided dataset evidence suggests that there is heterogeneity in inflammatory signaling within immune populations. Clusters 3, 18, 16, 2, and 8 are likely involved in inflammatory responses, as evidenced by the presence of specific genes involved in signaling pathways, matrix metalloproteinase regulation, cell adhesion, transcriptional regulation, and ribosomal protein functions. However, the exact mechanisms and relationships between these clusters and inflammatory signaling require further investigation. **Conclusion:** The provided dataset evidence suggests that there is heterogeneity in inflammatory signaling within immune populations, with different clusters exhibiting distinct gene expression profiles and potential involvement in inflammatory responses. Further analysis is required to elucidate the specific mechanisms and relationships between these clusters and inflammatory signaling.

Hybrid: Inflammatory signaling heterogeneity within immune populations

Based on the provided dataset evidence, I will analyze the clusters and provide a biological explanation for each. **Cluster 3 (n_cells = 667)** This cluster has a high expression similarity (0.9996904730796814) and a high hybrid similarity (0.8586544990539551). The metadata indicates that this cluster is predominantly composed of Myeloid cells, with a high proportion of Macrophages Monocyte-like cells (0.9460269865067467). The gene evidence shows high expression of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1, HSPB1) and other genes involved in cell stress response. This cluster may represent a population of macrophages that are highly responsive to inflammatory stimuli. **Cluster 18 (n_cells = 90)** This cluster has a moderate expression similarity (0.9996955394744873) and a moderate hybrid similarity (0.8523558974266052). The metadata indicates that this cluster is composed of Myeloid cells, with a high proportion of Mast cells (0.7888888888888889). The gene evidence shows high expression of genes involved in mast cell function, such

as ARHGAP15 and GATA2. This cluster may represent a population of mast cells that are involved in allergic responses. **Cluster 16** (n_cells = 141) This cluster has a high expression similarity (0.9998932480812073) and a moderate hybrid similarity (0.8512827157974243). The metadata indicates that this cluster is predominantly composed of Myeloid cells, with a high proportion of Macrophages-iron metabolism cells (0.9645390070921985). The gene evidence shows high expression of genes involved in iron metabolism, such as GPNMB and NEAT1. This cluster may represent a population of macrophages that are specialized in iron metabolism. **Cluster 2** (n_cells = 958) This cluster has a high expression similarity (0.9988120198249817) and a moderate hybrid similarity (0.8461151719093323). The metadata indicates that this cluster is predominantly composed of Myeloid cells, with a high proportion of Macrophages-iron metabolism cells (0.8507306889352818). The gene evidence shows high expression of heat shock proteins (HSPA1A, HSPA1B, HSPB1) and other genes involved in cell stress response. This cluster may represent a population of macrophages that are highly responsive to inflammatory stimuli. **Cluster 8** (n_cells = 342) This cluster has a high expression similarity (0.9997087717056274) and a moderate hybrid similarity (0.8442889451980591). The metadata indicates that this cluster is predominantly composed of Myeloid cells, with a high proportion of Monocytes (0.8216374269005848). The gene evidence shows high expression of genes involved in monocyte function, such as RPS26 and C1QC. This cluster may represent a population of monocytes that are involved in inflammatory responses. In summary, these clusters represent different populations of Myeloid cells that are specialized in various functions, including macrophage function, mast cell function, iron metabolism, and monocyte function. The gene evidence suggests that these clusters are characterized by high expression of specific genes involved in these functions.

Based on the provided dataset evidence, I will analyze the clusters and provide a biological explanation for each. **Cluster 16:** This cluster has a high hybrid similarity of 0.9998932480812073 and is primarily composed of macrophages, with a high expression of genes related to iron metabolism. The cluster is also enriched for cells from the "Meninges" dissection site. The genes GPNMB, NEAT1, and KLF6 are highly expressed in this cluster, suggesting a role in inflammatory signaling and cellular differentiation. **Cluster 17:** This cluster has a high hybrid similarity of 0.9998564720153809 and is primarily composed of lymphoid cells, specifically CD56-bright NK cells and ILC precursors. The cluster is also enriched for cells from the "Meninges" dissection site. The genes FYN, BTG1, and CD247 are highly expressed in this cluster, suggesting a role in immune cell development and function. **Cluster 19:** This cluster has a high hybrid similarity of 0.9997719526290894 and is primarily composed of myeloid cells, specifically macrophages and monocytes. The cluster is also enriched for cells from the "Meninges" dissection site. The genes RUNX1, LYVE1, and FOXO3 are highly expressed in this cluster, suggesting a role in cellular differentiation and immune function. **Cluster 8:** This cluster has a high hybrid similarity of 0.9997087717056274 and is primarily composed of myeloid cells, specifically monocytes and macrophages. The cluster is also enriched for cells from the "Meninges" dissection site. The genes RPS26, C1QC, and DAB2 are highly expressed in this cluster, suggesting a role in cellular differentiation and immune function. **Cluster 18:** This cluster has a high hybrid similarity of 0.9996955394744873 and is primarily composed of myeloid cells, specifically mast cells and monocytes. The cluster is also enriched for cells from the "Meninges" dissection site. The genes ARHGAP15, GATA2, and TIMP3 are highly expressed in this cluster, suggesting a role in cellular differentiation and immune function. In summary, these clusters are primarily composed of immune cells, specifically macrophages, lymphoid cells, and myeloid cells, and are enriched for cells from the "Meninges" dissection site. The highly expressed genes in each cluster suggest a role in inflammatory signaling, cellular differentiation, and immune function.

C.4.1 DISCOVERMODE DATASET 2

Metabolic stress programs co-opted by immune signaling pathways

****DATASET EVIDENCE**** Cluster 3: - HSPA1B (pct_in: 0.9940029985007496, pct_out: 0.523222992045163) - HSPA1A (pct_in: 0.9895052473763118, pct_out: 0.5243777264562484) - SLC9A9 (pct_in: 0.9760119940029985, pct_out: 0.5647934308442392) - MRC1 (pct_in: 0.9625187406296851, pct_out: 0.6203489863997947) - HSP90AA1 (pct_in: 0.9580209895052474, pct_out: 0.4638183217859892) - RPS4Y1 (pct_in: 0.9430284857571214, pct_out: 0.29907621247113164) - DNAJB1 (pct_in: 0.9415292353823088, pct_out: 0.5050038491147036) - STAB1 (pct_in: 0.9415292353823088, pct_out: 0.550423402617398) - HSPB1 (pct_in: 0.9340329835082459, pct_out: 0.48434693353861946) - JUN (pct_in: 0.9265367316341829, pct_out: 0.5282268411598665) Cluster 6: - HBA1 (pct_in: 0.981081081081081, pct_out: 0.42911877394636017) - HBA2 (pct_in: 0.981081081081081, pct_out: 0.42911877394636017) - HBZ (pct_in: 0.9675675675675676, pct_out: 0.22333456927450254) - ENSG00000284931 (pct_in: 0.9621621621621622, pct_out: 0.478556420714374) - ENSG00000239920 (pct_in: 0.9540540540540541, pct_out: 0.22172784575454207) - HPGDS (pct_in: 0.8081081081081081, pct_out: 0.5925101965146459) - C1QC (pct_in: 0.8027027027027027, pct_out: 0.6192065257693734) - C1QB (pct_in: 0.7891891891891892, pct_out: 0.5876900259547646) - FTL (pct_in: 0.7837837837837838, pct_out: 0.5836114200964034) - XIST (pct_in: 0.754054054054054, pct_out: 0.4592757384748486) Cluster 18: - ARHGAP15 (pct_in: 0.9666666666666667, pct_out: 0.4385378091028551) - GATA2 (pct_in: 0.9333333333333333, pct_out: 0.01290168438657269) - TIMP3 (pct_in: 0.8888888888888888, pct_out: 0.12519412256600168) - S100A6 (pct_in: 0.8888888888888888, pct_out: 0.40186357663361605) - GIHCG (pct_in: 0.8666666666666667, pct_out: 0.09246207143710429) - SLC24A3 (pct_in: 0.8444444444444444, pct_out: 0.025564448691912555) - MEIS2 (pct_in: 0.8333333333333334, pct_out: 0.02030820690479035) - ITGA4 (pct_in: 0.8222222222222222, pct_out: 0.1116951379763469) - TAGLN2 (pct_in: 0.8111111111111111, pct_out: 0.219687014693585) - SOX4 (pct_in: 0.8111111111111111, pct_out: 0.2956636005256242) Cluster 2: - RPS4Y1 (pct_in: 0.9749478079331941, pct_out: 0.2700253232040517) - HSPA1A (pct_in: 0.9665970772442589, pct_out: 0.5092629614820738) - HSPA1B (pct_in: 0.9634655532359081, pct_out: 0.5088631214180994) - SELENOP (pct_in: 0.9603340292275574, pct_out: 0.6085565773690523) - HSPB1 (pct_in: 0.9321503131524008, pct_out: 0.46714647474343596) - COLEC12 (pct_in: 0.9321503131524008, pct_out: 0.5800346528055444) - KIF5C-AS1 (pct_in: 0.9196242171189979, pct_out: 0.27828868452618954) - PMP22 (pct_in: 0.9164926931106472, pct_out: 0.524723443955751) - ASAH1 (pct_in: 0.9164926931106472, pct_out: 0.531520725043316) - LINC00278 (pct_in: 0.9154488517745303, pct_out: 0.21138211382113822) Cluster 16: - GPNMB (pct_in: 0.9858156028368794, pct_out: 0.24783653846153847) - NEAT1 (pct_in: 0.9787234042553191, pct_out: 0.5581730769230769) - SLC2A1 (pct_in: 0.9716312056737588, pct_out: 0.0829326923076923) - KLF6 (pct_in: 0.9645390070921985, pct_out: 0.5122596153846154) - FOS (pct_in: 0.9574468085106383, pct_out: 0.5524038461538462) - DUSP1 (pct_in: 0.9574468085106383, pct_out: 0.5461538461538461) - ATF3 (pct_in: 0.950354609929078, pct_out: 0.45252403846153844) - SH3BP5 (pct_in: 0.950354609929078, pct_out: 0.47872596153846153) - ARL4C (pct_in: 0.9432624113475178, pct_out: 0.4364182692307692) - SAMD4A (pct_in: 0.9148936170212766, pct_out: 0.40865384615384615) ****ESTABLISHED BIOLOGY**** - HSPA1B, HSPA1A: Heat shock proteins involved in protein folding and stress response in various cell types, including immune cells. - SLC9A9: Solute carrier family 9 member 9, involved in ion transport and pH regulation in immune cells. - MRC1: Mannose receptor C-type 1, involved in antigen presentation and immune response in macrophages. - HSP90AA1: Heat shock protein 90 alpha family class A member 1, involved in protein folding and stress response in various cell types, including immune cells. - RPS4Y1: Ribosomal protein S4 Y-linked, involved in protein synthesis and regulation in immune cells. - DNAJB1: DnaJ heat shock protein family

(Hsp40) member B1, involved in protein folding and stress response in various cell types, including immune cells. - STAB1: Stabilin 1, involved in endocytosis and immune response in macrophages. - HSPB1: Heat shock protein beta 1, involved in protein folding and stress response in various cell types, including immune cells. - JUN: Jun proto-oncogene, AP-1 transcription factor subunit, involved in cell proliferation and differentiation in immune cells. - HBA1, HBA2: Hemoglobin subunit alpha 1 and 2, involved in oxygen transport and immune response in red blood cells. - HBZ: HBZ protein, involved in immune response and regulation in T cells. - ENSG00000284931: Long non-coding RNA, involved in regulation of gene expression in immune cells. - ENSG00000239920: Long non-coding RNA, involved in regulation of gene expression in immune cells. - HPGDS: Hematopoietic prostaglandin D synthase, involved in immune response and regulation in T cells. - C1QC, C1QB: Complement component 1 q and b, involved in immune response and regulation in immune cells. - FTL: Ferritin light chain, involved in iron storage and regulation in immune cells. - XIST: X-inactive specific transcript, involved in X-chromosome inactivation and regulation in immune cells. - ARHGAP15: Rho GTPase activating protein 15, involved in regulation of cell migration and immune response in immune cells. - GATA2: GATA binding protein 2, involved in regulation of gene expression and immune response in immune cells. - TIMP3: Tissue inhibitor of metalloproteinase 3, involved in regulation of matrix metalloproteinases and immune response in immune cells. - S100A6: S100 calcium-binding protein A6, involved in regulation of immune response and cell proliferation in immune cells. - GIHCG: Glycoprotein hormone alpha chain, involved in regulation of immune response and cell proliferation in immune cells. - SLC24A3: Solute carrier family 24 member 3, involved in regulation of ion transport and immune response in immune cells. - MEIS2: Meis homeobox 2, involved in regulation of gene expression and immune response in immune cells. - ITGA4: Integrin subunit alpha 4, involved in regulation of cell adhesion and immune response in immune cells. - TAGLN2: Transgelin 2, involved in regulation of cell migration and immune response in immune cells. - SOX4: SRY-box transcription factor 4, involved in regulation of gene expression and immune response in immune cells. - RPS4Y1: Ribosomal protein S4 Y-linked, involved in protein synthesis and regulation in immune cells. - HSPA1A, HSPA1B: Heat shock proteins involved in protein folding and stress response in various cell types, including immune cells. - SELENOP: Selenoprotein P, involved in regulation of selenium metabolism and immune response in immune cells. - COLEC12: C-type lectin domain family 12 member A, involved in regulation of immune response and cell proliferation in immune cells. - KIF5C-AS1: Kinesin family member 5C antisense RNA 1, involved in regulation of gene expression and immune response in immune cells. - PMP22: Peripheral myelin protein 22, involved in regulation of myelination and immune response in immune cells. - ASAH1: N-acylsphingosine amidohydrolase 1, involved in regulation of sphingolipid metabolism and immune response in immune cells. - LINC00278: Long non-coding RNA, involved in regulation of gene expression and immune response in immune cells. - GPNMB: Glycoprotein NMB, involved in regulation of immune response and cell proliferation in immune cells. - NEAT1: Nuclear paraspeckle assembly transcript 1, involved in regulation of gene expression and immune response in immune cells. - SLC2A1: Solute carrier family 2 member 1, involved in regulation of glucose transport and immune response in immune cells. - KLF6: Kruppel-like factor 6, involved in regulation of gene expression and immune response in immune cells. - FOS: Fos proto-oncogene, AP-1 transcription factor subunit, involved in cell proliferation and differentiation in immune cells. - DUSP1: Dual specificity phosphatase 1, involved in regulation of MAP kinase signaling and immune response in immune cells. - ATF3: Activating transcription factor 3, involved in regulation of gene expression and immune response in immune cells. - SH3BP5: SH3 domain binding protein 5, involved in regulation of immune response and cell proliferation in immune cells. - ARL4C: ADP-ribosylation factor like GTPase 4C, involved in regulation of immune response and cell proliferation in immune cells. - SAMD4A: SAM domain, SH3 domain and nuclear localization signals 4A, involved in regulation of gene expression and immune response in immune cells. ****CONSISTENCY ANALYSIS** MATCHES ESTABLISHED BIOLOGY:** - HSPA1B, HSPA1A, HSP90AA1, HSPB1, JUN, RPS4Y1, DNAJB1, STAB1, HBA1, HBA2, HBZ, ENSG00000284931, ENSG00000239920, HPGDS, C1QC, C1QB, FTL, XIST, ARHGAP15, GATA2, TIMP3, S100A6, GIHCG, SLC24A3,

MEIS2, ITGA4, TAGLN2, SOX4, SELENOP, COLEC12, KIF5C-AS1, PMP22, ASAH1, LINC00278, GPNMB, NEAT1, SLC2A1, KLF6, FOS, DUSP1, ATF3, SH3BP5, ARL4C, SAMD4A UNEXPECTED/CONTEXT-SHIFTED: - SLC9A9: While involved in ion transport, its expression in immune cells is unexpected. - MRC1: While involved in antigen presentation, its expression in macrophages is expected but its high expression in cluster 3 is unexpected. - RPS4Y1: While involved in protein synthesis, its high expression in cluster 2 is unexpected. - ENSG00000284931, ENSG00000239920: While involved in regulation of gene expression, their expression in immune cells is unexpected. - HPGDS: While involved in immune response, its expression in cluster 6 is unexpected. - C1QC, C1QB: While involved in immune response, their expression in cluster 6 is unexpected. - FTL: While involved in iron storage, its expression in cluster 6 is unexpected. - XIST: While involved in X-chromosome inactivation, its expression in immune cells is unexpected. - GIHCG: While involved in immune response, its expression in cluster 18 is unexpected. - SLC24A3: While involved in ion transport, its expression in cluster 18 is unexpected. - MEIS2: While involved in regulation of gene expression, its expression in cluster 18 is unexpected. - ITGA4: While involved in cell adhesion, its expression in cluster 18 is unexpected. - TAGLN2: While involved in cell migration, its expression in cluster 18 is unexpected. - SOX4: While involved in regulation of gene expression, its expression in cluster 18 is unexpected. - KIF5C-AS1: While involved in regulation of gene expression, its expression in cluster 2 is unexpected. - PMP22: While involved in myelination, its expression in cluster 2 is unexpected. - ASAH1: While involved in sphingolipid metabolism, its expression in cluster 2 is unexpected. - LINC00278: While involved in regulation of gene expression, its expression in cluster 2 is unexpected. - NEAT1: While involved in regulation of gene expression, its expression in cluster 16 is unexpected. - SLC2A1: While involved in glucose transport, its expression in cluster 16 is unexpected. - KLF6: While involved in regulation of gene expression, its expression in cluster 16 is unexpected. - DUSP1: While involved in MAP kinase signaling, its expression in cluster 16 is unexpected. - ATF3: While involved in regulation of gene expression, its expression in cluster 16 is unexpected. - SH3BP5: While involved in immune response, its expression in cluster 16 is unexpected. - ARL4C: While involved in immune response, its expression in cluster 16 is unexpected. - SAMD4A: While involved in regulation of gene expression, its expression in cluster 16 is unexpected. ****CANDIDATE NOVEL HYPOTHESES**** - It is possible that SLC9A9 plays a role in ion transport in immune cells, but its expression in cluster 3 is unexpected and warrants further investigation. - The high expression of MRC1 in cluster 3 may indicate a novel role in antigen presentation in macrophages. - The high expression of RPS4Y1 in cluster 2 may indicate a novel role in protein synthesis in immune cells. - The expression of ENSG00000284931 and ENSG00000239920 in immune cells may indicate a novel role in regulation of gene expression. - The expression of HPGDS, C1QC, C1QB, FTL, and XIST in cluster 6 may indicate a novel role in immune response and regulation in immune cells. - The expression of GIHCG, SLC24A3, MEIS2, ITGA4, TAGLN2, SOX4, KIF5C-AS1, PMP22, ASAH1, LINC00278, NEAT1, SLC2A1, KLF6, DUSP1, ATF3, SH3BP5, ARL4C, and SAMD4A in cluster 18 may indicate a novel role in immune response and regulation in immune cells. - The expression of these genes in specific clusters may indicate a novel role in metabolic stress programs co-opted by immune signaling pathways.

Structural and cytoskeletal gene programs associated with immune regulation

****1. DATASET EVIDENCE**** Cluster 3: - HSPA1B (pct.in: 0.9940029985007496) - HSPA1A (pct.in: 0.9895052473763118) - SLC9A9 (pct.in: 0.9760119940029985) - MRC1 (pct.in: 0.9625187406296851) - HSP90AA1 (pct.in: 0.9580209895052474) - RPS4Y1 (pct.in: 0.9430284857571214) - DNAJB1 (pct.in: 0.9415292353823088) - STAB1 (pct.in: 0.9415292353823088) - HSPB1 (pct.in: 0.9340329835082459) - JUN (pct.in: 0.9265367316341829) Cluster 13: - TTR (pct.in: 0.9878787878787879) - CD74 (pct.in: 0.9818181818181818) - PLXDC2 (pct.in: 0.9818181818181818) - RPL41 (pct.in: 0.9696969696969697) - FOS (pct.in: 0.9696969696969697) - IER2 (pct.in: 0.9636363636363636) - S100A11 (pct.in: 0.9575757575757575) - TEX14 (pct.in: 0.9515151515151515) - RPL39 (pct.in: 0.9393939393939394) - ALOX5AP

(pct.in: 0.9272727272727272) Cluster 11: - PLXDC2 (pct.in: 0.9629629629629629)
 - SLC8A1 (pct.in: 0.9037037037037037) - S100A11 (pct.in: 0.8703703703703703)
 - FRMD4A (pct.in: 0.8555555555555555) - DOCK4 (pct.in: 0.8481481481481481)
 - TREM2 (pct.in: 0.837037037037037) - ALOX5AP (pct.in: 0.825925925925926) -
 CAPG (pct.in: 0.8222222222222222) - LIMS1 (pct.in: 0.8185185185185185) - C3
 (pct.in: 0.7888888888888889) Cluster 18: - ARHGAP15 (pct.in: 0.9666666666666667)
 - GATA2 (pct.in: 0.9333333333333333) - TIMP3 (pct.in: 0.8888888888888888) -
 S100A6 (pct.in: 0.8888888888888888) - GIHCG (pct.in: 0.8666666666666667) -
 SLC24A3 (pct.in: 0.8444444444444444) - MEIS2 (pct.in: 0.8333333333333334) -
 ITGA4 (pct.in: 0.8222222222222222) - TAGLN2 (pct.in: 0.8111111111111111) -
 SOX4 (pct.in: 0.8111111111111111) Cluster 7: - CD74 (pct.in: 0.9564032697547684)
 - VIM (pct.in: 0.8937329700272479) - HLA-DRA (pct.in: 0.8637602179836512) -
 CST3 (pct.in: 0.8419618528610354) - TIMP1 (pct.in: 0.7983651226158038) - S100A4
 (pct.in: 0.7929155313351499) - HLA-DRB1 (pct.in: 0.779291553133515) - HLA-DPA1
 (pct.in: 0.7520435967302452) - MNDA (pct.in: 0.7356948228882834) - ANXA1 (pct.in:
 0.7275204359673024) **2. ESTABLISHED BIOLOGY** - HSPA1B, HSPA1A: Heat
 shock proteins involved in protein folding and stress response. - SLC9A9: Solute carrier
 family 9 member 9, involved in ion transport and cellular homeostasis. - MRC1: Mannose
 receptor C type 1, involved in antigen presentation and immune response. - HSP90AA1:
 Heat shock protein 90 alpha family class A member 1, involved in protein folding and signal
 transduction. - RPS4Y1: Ribosomal protein S4 Y-linked, involved in protein synthesis. -
 DNAJB1: DnaJ heat shock protein family (Hsp40) member B1, involved in protein folding
 and stress response. - STAB1: Stabilin 1, involved in endocytosis and immune response. -
 HSPB1: Heat shock protein beta 1, involved in protein folding and stress response. - JUN:
 Jun proto-oncogene, involved in cell growth and differentiation. - TTR: Transthyretin,
 involved in protein transport and metabolism. - CD74: CD74 molecule, involved in antigen
 presentation and immune response. - PLXDC2: Phospholipid scramblase 3, involved in
 phospholipid metabolism and immune response. - RPL41: Ribosomal protein L41, involved
 in protein synthesis. - FOS: Fos proto-oncogene, involved in cell growth and differentia-
 tion. - IER2: Immediate early response 2, involved in cell growth and differentiation. -
 S100A11: S100 calcium-binding protein A11, involved in calcium signaling and immune
 response. - TEX14: Testis expressed 14, involved in spermatogenesis and immune response.
 - RPL39: Ribosomal protein L39, involved in protein synthesis. - ALOX5AP: Arachidonate
 5-lipoxygenase activating protein, involved in leukotriene synthesis and immune response.
 - PLXDC2: Phospholipid scramblase 3, involved in phospholipid metabolism and immune
 response. - SLC8A1: Solute carrier family 8 member A1, involved in ion transport and
 cellular homeostasis. - FRMD4A: FERM domain containing 4A, involved in cell adhesion
 and immune response. - DOCK4: Deducator of cytokinesis 4, involved in cell signaling and
 immune response. - TREM2: Triggering receptor expressed on myeloid cells 2, involved in
 immune response and neurodegeneration. - ALOX5AP: Arachidonate 5-lipoxygenase acti-
 vating protein, involved in leukotriene synthesis and immune response. - CAPG: Capping
 protein (actin filament), involved in actin cytoskeleton organization and immune response.
 - LIMS1: LIM and senescent cell antigen-like containing 1, involved in cell adhesion and
 immune response. - C3: Complement component 3, involved in immune response and in-
 flammation. - ARHGAP15: Rho GTPase activating protein 15, involved in cell signaling
 and immune response. - GATA2: GATA binding protein 2, involved in cell growth and
 differentiation. - TIMP3: Tissue inhibitor of metalloproteinase 3, involved in matrix metal-
 loproteinase regulation and immune response. - S100A6: S100 calcium-binding protein A6,
 involved in calcium signaling and immune response. - GIHCG: Glycoprotein hormone alpha
 polypeptide, involved in cell growth and differentiation. - SLC24A3: Solute carrier family
 24 member 3, involved in ion transport and cellular homeostasis. - MEIS2: Meis homeobox
 2, involved in cell growth and differentiation. - ITGA4: Integrin subunit alpha 4, involved
 in cell adhesion and immune response. - TAGLN2: Transgelin 2, involved in actin cytoskeleton
 organization and immune response. - SOX4: SRY-box transcription factor 4, involved in cell
 growth and differentiation. - CD74: CD74 molecule, involved in antigen presentation and
 immune response. - VIM: Vimentin, involved in actin cytoskeleton organization and immune
 response. - HLA-DRA: HLA class II histocompatibility antigen, DR alpha chain, involved in

antigen presentation and immune response. - CST3: Cystatin C, involved in cysteine protease regulation and immune response. - TIMP1: Tissue inhibitor of metalloproteinase 1, involved in matrix metalloproteinase regulation and immune response. - S100A4: S100 calcium-binding protein A4, involved in calcium signaling and immune response. - HLA-DRB1: HLA class II histocompatibility antigen, DR beta 1 chain, involved in antigen presentation and immune response. - HLA-DPA1: HLA class II histocompatibility antigen, DP alpha 1 chain, involved in antigen presentation and immune response. - MND1: Myeloid cell nuclear differentiation antigen, involved in cell growth and differentiation. - ANXA1: Annexin A1, involved in phospholipid metabolism and immune response. **3. CONSISTENCY ANALYSIS** MATCHES ESTABLISHED BIOLOGY: - HSPA1B, HSPA1A, HSP90AA1, HSPB1, JUN, CD74, PLXDC2, RPL41, FOS, IER2, S100A11, TEX14, RPL39, ALOX5AP, CAPG, LIMS1, C3, ARHGAP15, GATA2, TIMP3, S100A6, GIHCG, SLC24A3, MEIS2, ITGA4, TAGLN2, SOX4, CD74, VIM, HLA-DRA, CST3, TIMP1, S100A4, HLA-DRB1, HLA-DPA1, MND1, ANXA1: These genes are known to be involved in various biological processes, including immune response, cell growth and differentiation, and protein folding. UNEXPECTED OR CONTEXT-SHIFTED: - SLC9A9: While involved in ion transport and cellular homeostasis, its expression in immune cells is unexpected. - MRC1: While involved in antigen presentation and immune response, its expression in immune cells is unexpected. - PLXDC2: While involved in phospholipid metabolism and immune response, its expression in immune cells is unexpected. - FRMD4A: While involved in cell adhesion and immune response, its expression in immune cells is unexpected. - DOCK4: While involved in cell signaling and immune response, its expression in immune cells is unexpected. - TREM2: While involved in immune response and neurodegeneration, its expression in immune cells is unexpected. - SLC8A1: While involved in ion transport and cellular homeostasis, its expression in immune cells is unexpected. - S100A6: While involved in calcium signaling and immune response, its expression in immune cells is unexpected. - GIHCG: While involved in cell growth and differentiation, its expression in immune cells is unexpected. - SLC24A3: While involved in ion transport and cellular homeostasis, its expression in immune cells is unexpected. - MEIS2: While involved in cell growth and differentiation, its expression in immune cells is unexpected. - TAGLN2: While involved in actin cytoskeleton organization and immune response, its expression in immune cells is unexpected. - SOX4: While involved in cell growth and differentiation, its expression in immune cells is unexpected. **4. CANDIDATE NOVEL HYPOTHESES** - It is possible that SLC9A9 plays a role in modulating ion transport and cellular homeostasis in immune cells, potentially influencing immune response. - MRC1 may be involved in antigen presentation and immune response in immune cells, potentially influencing immune response. - PLXDC2 may be involved in phospholipid metabolism and immune response in immune cells, potentially influencing immune response. - FRMD4A may be involved in cell adhesion and immune response in immune cells, potentially influencing immune response. - DOCK4 may be involved in cell signaling and immune response in immune cells, potentially influencing immune response. - TREM2 may be involved in immune response and neurodegeneration in immune cells, potentially influencing immune response. - SLC8A1 may be involved in ion transport and cellular homeostasis in immune cells, potentially influencing immune response. - S100A6 may be involved in calcium signaling and immune response in immune cells, potentially influencing immune response. - GIHCG may be involved in cell growth and differentiation in immune cells, potentially influencing immune response. - SLC24A3 may be involved in ion transport and cellular homeostasis in immune cells, potentially influencing immune response. - MEIS2 may be involved in cell growth and differentiation in immune cells, potentially influencing immune response. - TAGLN2 may be involved in actin cytoskeleton organization and immune response in immune cells, potentially influencing immune response. - SOX4 may be involved in cell growth and differentiation in immune cells, potentially influencing immune response. Note: These hypotheses are based on the unexpected or context-shifted expression of these genes in immune cells and are not necessarily causal. Further investigation is needed to confirm these hypotheses.

Reactivation of developmental transcriptional programs during immune responses

****1. DATASET EVIDENCE**** - ****Cluster 3:**** - HSPA1B (pct.in: 0.9940029985, pct.out: 0.523222992) - HSPA1A (pct.in: 0.9895052474, pct.out: 0.524377726) - SLC9A9 (pct.in: 0.976011994, pct.out: 0.564793430) - MRC1 (pct.in: 0.9625187406, pct.out: 0.620348986) - HSP90AA1 (pct.in: 0.9580209895, pct.out: 0.463818321) - RPS4Y1 (pct.in: 0.9430284858, pct.out: 0.299076212) - DNAJB1 (pct.in: 0.9415292354, pct.out: 0.505003849) - STAB1 (pct.in: 0.9415292354, pct.out: 0.550423402) - HSPB1 (pct.in: 0.9340329835, pct.out: 0.484346933) - JUN (pct.in: 0.9265367316, pct.out: 0.528226841) - ****Cluster 18:**** - ARHGAP15 (pct.in: 0.9666666667, pct.out: 0.438537809) - GATA2 (pct.in: 0.9333333333, pct.out: 0.012901684) - TIMP3 (pct.in: 0.8888888889, pct.out: 0.125194122) - S100A6 (pct.in: 0.8888888889, pct.out: 0.401863576) - GIHCG (pct.in: 0.8666666667, pct.out: 0.092462071) - SLC24A3 (pct.in: 0.8444444444, pct.out: 0.025564449) - MEIS2 (pct.in: 0.8333333333, pct.out: 0.020308207) - ITGA4 (pct.in: 0.8222222222, pct.out: 0.111695138) - TAGLN2 (pct.in: 0.8111111111, pct.out: 0.219687015) - SOX4 (pct.in: 0.8111111111, pct.out: 0.295663600) - ****Cluster 1:**** - LGMN (pct.in: 0.8996789727, pct.out: 0.610533610) - C1QA (pct.in: 0.8924558587, pct.out: 0.617463617) - DAB2 (pct.in: 0.8836276083, pct.out: 0.573388773) - FCGRT (pct.in: 0.8804173355, pct.out: 0.637560637) - BLVRB (pct.in: 0.8627608347, pct.out: 0.569369369) - PLTP (pct.in: 0.8451043339, pct.out: 0.505751906) - VSIG4 (pct.in: 0.8378812199, pct.out: 0.569230769) - A2M (pct.in: 0.834670947, pct.out: 0.565072765) - C1QB (pct.in: 0.8314606742, pct.out: 0.555925156) - CD14 (pct.in: 0.8298555377, pct.out: 0.597782398) - ****Cluster 16:**** - GPNMB (pct.in: 0.9858156028, pct.out: 0.247836538) - NEAT1 (pct.in: 0.9787234043, pct.out: 0.558173077) - SLC2A1 (pct.in: 0.9716312057, pct.out: 0.082932693) - KLF6 (pct.in: 0.9645390071, pct.out: 0.512259615) - FOS (pct.in: 0.9574468085, pct.out: 0.552403846) - DUSP1 (pct.in: 0.9574468085, pct.out: 0.546153846) - ATF3 (pct.in: 0.9503546099, pct.out: 0.452524038) - SH3BP5 (pct.in: 0.9503546099, pct.out: 0.478725961) - ARL4C (pct.in: 0.9432624113, pct.out: 0.436418269) - SAMD4A (pct.in: 0.914893617, pct.out: 0.408653846) - ****Cluster 8:**** - RPS26 (pct.in: 0.9590643275, pct.out: 0.467052592) - C1QC (pct.in: 0.8947368421, pct.out: 0.615962557) - RPL39 (pct.in: 0.8801169591, pct.out: 0.469392782) - DAB2 (pct.in: 0.8684210526, pct.out: 0.608572484) - CST3 (pct.in: 0.865497076, pct.out: 0.569035596) - AGR2 (pct.in: 0.8391812865, pct.out: 0.502401774) - RNASE1 (pct.in: 0.8304093567, pct.out: 0.697130188) - RPL41 (pct.in: 0.8099415205, pct.out: 0.507574824) - FTL (pct.in: 0.8070175439, pct.out: 0.583323069) - BLVRB (pct.in: 0.8040935673, pct.out: 0.604507944) ****2. ESTABLISHED BIOLOGY**** - ****HSPA1B, HSPA1A, HSP90AA1, HSPB1:**** Heat shock proteins are involved in protein folding and stress response. They are often upregulated in response to cellular stress. - ****SLC9A9, MRC1:**** SLC9A9 is involved in the regulation of intracellular pH, while MRC1 is involved in the regulation of immune responses. - ****RPS4Y1:**** RPS4Y1 is a ribosomal protein involved in protein synthesis. - ****DNAJB1, STAB1:**** DNAJB1 is involved in protein folding and degradation, while STAB1 is involved in the regulation of cell adhesion. - ****JUN:**** JUN is a transcription factor involved in the regulation of cell growth and differentiation. - ****ARHGAP15:**** ARHGAP15 is involved in the regulation of Rho GTPase activity, which is important for cell migration and cytoskeleton organization. - ****GATA2:**** GATA2 is a transcription factor involved in the regulation of hematopoiesis and immune responses. - ****TIMP3, S100A6:**** TIMP3 is involved in the regulation of matrix metalloproteinase activity, while S100A6 is involved in the regulation of cell proliferation and differentiation. - ****GIHCG:**** GIHCG is involved in the regulation of cell growth and differentiation. - ****SLC24A3:**** SLC24A3 is involved in the regulation of cell adhesion and migration. - ****MEIS2:**** MEIS2 is a transcription factor involved in the regulation of hematopoiesis and immune responses. - ****ITGA4:**** ITGA4 is involved in the regulation of cell adhesion and migration. - ****TAGLN2, SOX4:**** TAGLN2 is involved in the regulation of cell adhesion and migration, while SOX4 is a transcription factor involved in the regulation of cell growth and differentiation. - ****LGMN:**** LGMN is involved in the regulation of cell growth and differentiation. - ****C1QA, C1QB:**** C1QA and C1QB are involved in the regulation of immune responses. - ****DAB2:**** DAB2 is involved in the regulation of cell growth and differentiation. - ****FCGRT:**** FCGRT is involved in the regulation of immune responses. - ****BLVRB:**** BLVRB is involved in the regulation of cell growth

and differentiation. - **PLTP**: PLTP is involved in the regulation of lipid metabolism. - **VSIG4**: VSIG4 is involved in the regulation of immune responses. - **A2M**: A2M is involved in the regulation of immune responses. - **CD14**: CD14 is involved in the regulation of immune responses. - **GPNMB**: GPNMB is involved in the regulation of cell growth and differentiation. - **NEAT1**: NEAT1 is involved in the regulation of cell growth and differentiation. - **SLC2A1**: SLC2A1 is involved in the regulation of glucose metabolism. - **KLF6**: KLF6 is a transcription factor involved in the regulation of cell growth and differentiation. - **FOS**: FOS is a transcription factor involved in the regulation of cell growth and differentiation. - **DUSP1**: DUSP1 is involved in the regulation of MAP kinase activity. - **ATF3**: ATF3 is a transcription factor involved in the regulation of cell growth and differentiation. - **SH3BP5**: SH3BP5 is involved in the regulation of cell growth and differentiation. - **ARL4C**: ARL4C is involved in the regulation of cell growth and differentiation. - **SAMD4A**: SAMD4A is involved in the regulation of cell growth and differentiation. - **RPS26**: RPS26 is a ribosomal protein involved in protein synthesis. - **C1QC**: C1QC is involved in the regulation of immune responses. - **RPL39**: RPL39 is a ribosomal protein involved in protein synthesis. - **CST3**: CST3 is involved in the regulation of cell growth and differentiation. - **AGR2**: AGR2 is involved in the regulation of cell growth and differentiation. - **RNASE1**: RNASE1 is involved in the regulation of cell growth and differentiation. - **RPL41**: RPL41 is a ribosomal protein involved in protein synthesis. - **FTL**: FTL is involved in the regulation of cell growth and differentiation. - **BLVRB**: BLVRB is involved in the regulation of cell growth and differentiation. **3. CONSISTENCY ANALYSIS** - **Expected observations**: - Heat shock proteins (HSPA1B, HSPA1A, HSP90AA1, HSPB1) are upregulated in response to cellular stress. - SLC9A9 and MRC1 are involved in the regulation of intracellular pH and immune responses, respectively. - RPS4Y1 is a ribosomal protein involved in protein synthesis. - DNAJB1 and STAB1 are involved in protein folding and degradation, and cell adhesion, respectively. - JUN is a transcription factor involved in the regulation of cell growth and differentiation. - ARHGAP15 is involved in the regulation of Rho GTPase activity. - GATA2 is a transcription factor involved in the regulation of hematopoiesis and immune responses. - TIMP3 and S100A6 are involved in the regulation of matrix metalloproteinase activity and cell proliferation and differentiation, respectively. - GIHCG is involved in the regulation of cell growth and differentiation. - SLC24A3 is involved in the regulation of cell adhesion and migration. - MEIS2 is a transcription factor involved in the regulation of hematopoiesis and immune responses. - ITGA4 is involved in the regulation of cell adhesion and migration. - TAGLN2 and SOX4 are involved in the regulation of cell adhesion and migration, and cell growth and differentiation, respectively. - LGMN is involved in the regulation of cell growth and differentiation. - C1QA and C1QB are involved in the regulation of immune responses. - DAB2 is involved in the regulation of cell growth and differentiation. - FCGRT is involved in the regulation of immune responses. - BLVRB is involved in the regulation of cell growth and differentiation. - PLTP is involved in the regulation of lipid metabolism. - VSIG4 is involved in the regulation of immune responses. - A2M is involved in the regulation of immune responses. - CD14 is involved in the regulation of immune responses. - GPNMB is involved in the regulation of cell growth and differentiation. - NEAT1 is involved in the regulation of cell growth and differentiation. - SLC2A1 is involved in the regulation of glucose metabolism. - KLF6 is a transcription factor involved in the regulation of cell growth and differentiation. - FOS is a transcription factor involved in the regulation of cell growth and differentiation. - DUSP1 is involved in the regulation of MAP kinase activity. - ATF3 is a transcription factor involved in the regulation of cell growth and differentiation. - SH3BP5 is involved in the regulation of cell growth and differentiation. - ARL4C is involved in the regulation of cell growth and differentiation. - SAMD4A is involved in the regulation of cell growth and differentiation. - RPS26 is a ribosomal protein involved in protein synthesis. - C1QC is involved in the regulation of immune responses. - RPL39 is a ribosomal protein involved in protein synthesis. - CST3 is involved in the regulation of cell growth and differentiation. - AGR2 is involved in the regulation of cell growth and differentiation. - RNASE1 is involved in the regulation of cell growth and differentiation. - RPL41 is a ribosomal protein involved in protein synthesis. - FTL is involved in the regulation of cell growth and differentiation. - BLVRB is involved in the reg-

ulation of cell growth and differentiation. - ****Unexpected or context-shifted observations:****

- The upregulation of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1, HSPB1) in response to immune responses is unexpected.
- The involvement of SLC9A9 and MRC1 in the regulation of intracellular pH and immune responses, respectively, is unexpected.
- The upregulation of RPS4Y1 in response to immune responses is unexpected.
- The involvement of DNAJB1 and STAB1 in protein folding and degradation, and cell adhesion, respectively, is unexpected.
- The upregulation of JUN in response to immune responses is unexpected.
- The involvement of ARHGAP15 in the regulation of Rho GTPase activity is unexpected.
- The upregulation of GATA2 in response to immune responses is unexpected.
- The involvement of TIMP3 and S100A6 in the regulation of matrix metalloproteinase activity and cell proliferation and differentiation, respectively, is unexpected.
- The upregulation of GIHCG in response to immune responses is unexpected.
- The involvement of SLC24A3 in the regulation of cell adhesion and migration is unexpected.
- The upregulation of MEIS2 in response to immune responses is unexpected.
- The involvement of ITGA4 in the regulation of cell adhesion and migration is unexpected.
- The upregulation of TAGLN2 and SOX4 in response to immune responses is unexpected.
- The involvement of LGMN in the regulation of cell growth and differentiation is unexpected.
- The upregulation of C1QA and C1QB in response to immune responses is unexpected.
- The involvement of DAB2 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of FCGRT in response to immune responses is unexpected.
- The involvement of BLVRB in the regulation of cell growth and differentiation is unexpected.
- The upregulation of PLTP in response to immune responses is unexpected.
- The involvement of VSIG4 in the regulation of immune responses is unexpected.
- The upregulation of A2M in response to immune responses is unexpected.
- The involvement of CD14 in the regulation of immune responses is unexpected.
- The upregulation of GPNMB in response to immune responses is unexpected.
- The involvement of NEAT1 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of SLC2A1 in response to immune responses is unexpected.
- The involvement of KLF6 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of FOS in response to immune responses is unexpected.
- The involvement of DUSP1 in the regulation of MAP kinase activity is unexpected.
- The upregulation of ATF3 in response to immune responses is unexpected.
- The involvement of SH3BP5 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of ARL4C in response to immune responses is unexpected.
- The involvement of SAMD4A in the regulation of cell growth and differentiation is unexpected.
- The upregulation of RPS26 in response to immune responses is unexpected.
- The involvement of C1QC in the regulation of immune responses is unexpected.
- The upregulation of RPL39 in response to immune responses is unexpected.
- The involvement of CST3 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of AGR2 in response to immune responses is unexpected.
- The involvement of RNASE1 in the regulation of cell growth and differentiation is unexpected.
- The upregulation of RPL41 in response to immune responses is unexpected.
- The involvement of FTL in the regulation of cell growth and differentiation is unexpected.
- The upregulation of BLVRB in response to immune responses is unexpected.

****4. CANDIDATE NOVEL HYPOTHESES**** - ****Probabilistic hypotheses:****

- The upregulation of heat shock proteins (HSPA1B, HSPA1A, HSP90AA1, HSPB1) in response to immune responses may be indicative of a stress response mechanism that is activated in response to immune stimuli.
- The involvement of SLC9A9 and MRC1 in the regulation of intracellular pH and immune responses, respectively, may indicate a novel mechanism of immune regulation that involves the regulation of intracellular pH.
- The upregulation of RPS4Y1 in response to immune responses may indicate a novel mechanism of protein synthesis that is activated in response to immune stimuli.
- The involvement of DNAJB1 and STAB1 in protein folding and degradation, and cell adhesion, respectively, may indicate a novel mechanism of protein regulation that is activated in response to immune stimuli.
- The upregulation of JUN in response to immune responses may indicate a novel mechanism of transcriptional regulation that is activated in response to immune stimuli.
- The involvement of ARHGAP15 in the regulation of Rho GTPase activity may indicate a novel mechanism of cell migration and cytoskeleton organization that is activated in response to immune stimuli.
- The upregulation of GATA2 in response to immune responses may indicate a novel mechanism of hematopoiesis and

immune regulation that is activated in response to immune stimuli. - The involvement of TIMP3 and S100A6 in the regulation of matrix metalloproteinase activity and cell proliferation and differentiation, respectively, may indicate a novel mechanism of immune regulation that involves the regulation of matrix metalloproteinase activity and cell proliferation and differentiation. - The upregulation of GIHCG in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli. - The involvement of SLC24A3 in the regulation of cell adhesion and migration may indicate a novel mechanism of immune regulation that involves the regulation of cell adhesion and migration. - The upregulation of MEIS2 in response to immune responses may indicate a novel mechanism of hematopoiesis and immune regulation that is activated in response to immune stimuli. - The involvement of ITGA4 in the regulation of cell adhesion and migration may indicate a novel mechanism of immune regulation that involves the regulation of cell adhesion and migration. - The upregulation of TAGLN2 and SOX4 in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli. - The involvement of LGMN in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of C1QA and C1QB in response to immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The involvement of DAB2 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of FCGRT in response to immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The involvement of BLVRB in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of PLTP in response to immune responses may indicate a novel mechanism of lipid metabolism that is activated in response to immune stimuli. - The involvement of VSIG4 in the regulation of immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The upregulation of A2M in response to immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The involvement of CD14 in the regulation of immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The upregulation of GPNMB in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli. - The involvement of NEAT1 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of SLC2A1 in response to immune responses may indicate a novel mechanism of glucose metabolism that is activated in response to immune stimuli. - The involvement of KLF6 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of FOS in response to immune responses may indicate a novel mechanism of transcriptional regulation that is activated in response to immune stimuli. - The involvement of DUSP1 in the regulation of MAP kinase activity may indicate a novel mechanism of immune regulation that involves the regulation of MAP kinase activity. - The upregulation of ATF3 in response to immune responses may indicate a novel mechanism of transcriptional regulation that is activated in response to immune stimuli. - The involvement of SH3BP5 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of ARL4C in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli. - The involvement of SAMD4A in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of RPS26 in response to immune responses may indicate a novel mechanism of protein synthesis that is activated in response to immune stimuli. - The involvement of C1QC in the regulation of immune responses may indicate a novel mechanism of immune regulation that involves the regulation of immune responses. - The upregulation of RPL39 in response to immune responses may indicate a novel mecha-

nism of protein synthesis that is activated in response to immune stimuli. - The involvement of CST3 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of AGR2 in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli. - The involvement of RNASE1 in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of RPL41 in response to immune responses may indicate a novel mechanism of protein synthesis that is activated in response to immune stimuli. - The involvement of FTL in the regulation of cell growth and differentiation may indicate a novel mechanism of immune regulation that involves the regulation of cell growth and differentiation. - The upregulation of BLVRB in response to immune responses may indicate a novel mechanism of cell growth and differentiation that is activated in response to immune stimuli.