ELISA: An Interpretable Hybrid Generative AI Agent for Expression-Grounded Discovery in Single-Cell Genomics

Omar Coser, looking for Postdoc position in AI+ Bioinformatics [1]

Abstract

Translating single-cell RNA sequencing (scRNA-seq) data into mechanistic biological hypotheses remains a critical bottleneck, as agentic AI systems lack direct access to transcriptomic representations while expression foundation models remain opaque to natural language. Here we introduce ELISA (Embedding-Linked Interactive Single-cell Agent), an interpretable framework that unifies scGPT expression embeddings with BioBERT-based semantic retrieval and LLM-mediated interpretation for interactive single-cell discovery. An automatic query classifier routes inputs to gene marker scoring, semantic matching, or reciprocal rank fusion pipelines depending on whether the query is a gene signature, natural language concept, or mixture of both. Integrated analytical modules perform pathway activity scoringacross 60+ gene sets, ligand–receptor interaction prediction using 280+ curated pairs, condition-aware comparative analysis, and cell-type proportion estimation all operating directly on embedded data without access to the original count matrix. Benchmarked across six diverse scRNA-seq datasets spanning inflammatory lung disease, pediatric and adult cancers, organoid models, healthy tissue, and neurodevelopment, ELISA significantly outperforms CellWhisperer in cell type retrieval (combined permutation test, $p < 0.001$), with particularly large gains on gene-signature queries (Cohen's $d = 5.98$ for MRR). ELISA replicates published biological findings (mean composite score 0.90) with near-perfect pathway alignment and theme coverage (0.98 each), and generates candidate hypotheses through grounded LLM reasoning, bridging the gap between transcriptomic data exploration and biological discovery.

[1]No affiliation. Correspondence to: Omar Coser <omarcoser10@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed our understanding of cellular heterogeneity by enabling genome-wide transcriptional profiling at single-cell resolution (Tang et al., 2009). Standardized pipelines for quality control, normalization, clustering, differential expression, and trajectory inference (Luecken & Theis, 2019) have catalyzed comprehensive cell atlases across tissues, developmental stages, and disease contexts. Yet a critical bottleneck persists: translating differentially expressed gene lists, enriched pathways, and predicted ligand–receptor interactions into mechanistic biological hypotheses remains labor-intensive, context-dependent, and difficult to scale.

Large-language models (LLMs) offer a potential solution. They encode substantial biomedical knowledge and perform competitively on clinical reasoning benchmarks (Singhal et al., 2023), while retrieval-augmented generation (RAG) improves factual accuracy by grounding outputs in external knowledge (Lewis et al., 2020). These capabilities have motivated *agentic* architectures capable of autonomous planning, tool use, and iterative reasoning.

Recent agentic systems span a broad range of biomedical applications (Table 1). **Towards an AI Co-Scientist** (Gottweis et al., 2025) introduces multi-agent hypothesis generation through structured debate, but operates over text without interfacing with experimental data. **Biomni** (Huang et al., 2025) unifies biomedical tools and databases into a single action space, enabling tasks such as gene prioritization. **GeneAgent** (Wang et al., 2025) and related work (Gao et al., 2024) extend LLM reasoning to gene-set analysis, while **Virtual Lab** (Swanson et al., 2025) demonstrates collaborative multi-agent discovery. Within single-cell analysis, **CellAgent** (Xiao et al., 2024) decomposes scRNA-seq workflows into agent-handled subtasks, **AutoBA** (Zhou et al., 2023) generates executable pipelines from natural language, and **BRAD** (Pickard et al., 2025) integrates LLMs with enrichment analysis for biomarker identification. For retrieval, **GeneGPT** (Jin et al., 2025) provides structured access to NCBI databases, and systems for deep phenotyping (Garcia et al., 2025) and biomedical extraction (Cinquin, 2024; Niyonkuru et al., 2025) demonstrate the utility of RAG. **CRISPR-GPT** (Qu et al., 2025) further illustrates agen-

tic automation for gene-editing design. However, these systems rely on curated text and structured databases and cannot operate directly on high-dimensional transcriptomic representations.

Concurrently, foundation models for single-cell biology have learned expressive latent representations from transcriptomic data. **scGPT** (Cui et al., 2024) uses generative pre-training over millions of single cells, capturing gene–gene dependencies for embedding, annotation transfer, and perturbation prediction, while extensions such as **scWGBS-GPT** (Liang et al., 2025) and **Tokensome** (Zhang et al., 2024) broaden these representations to methylomics and multimodal settings. However, expression embeddings are not designed for semantic querying: they capture transcriptional similarity in latent spaces unaligned with the natural-language concepts biologists use to formulate hypotheses. **CellWhisperer** (Schaefer et al., 2025) partly addresses this by learning joint embeddings of transcriptomes and textual annotations via contrastive training, enabling chat-based interrogation within CELLxGENE, but it lacks built-in modules for pathway scoring, interaction prediction, or condition-aware comparison.

This landscape reveals a fundamental disconnect: agentic systems excel at reasoning over text but lack direct access to transcriptional data structure, while expression foundation models learn rich cellular representations that remain opaque to natural-language interfaces. No existing system unifies expression-derived embeddings with semantic language representations in a single interactive framework for single-cell discovery.

ELISA (**E**mbedding-**L**inked **I**nteractive **S**ingle-cell **A**gent) addresses this gap by integrating scGPT expression embeddings with semantic retrieval and LLM-based interpretation in a unified discovery platform (Fig. 4). Rather than retraining foundation models, ELISA combines scGPT cluster embeddings with BioBERT-derived semantic embeddings through a hybrid routing mechanism. A query classifier detects whether the input is a gene signature, a natural-language concept, or a mixture, and routes it to gene marker scoring, semantic cosine similarity, or reciprocal rank fusion of both. Built-in modules for condition-aware comparison, ligand-receptor prediction, pathway scoring, and cell-type proportion analysis operate directly on the embedded data, while an LLM reasoning layer translates statistical outputs into structured biological interpretations. Critically, ELISA enforces strict separation between dataset-derived evidence and LLM-generated knowledge, supporting transparent hypothesis generation and producing publication-ready reports with Nature-style visualizations.

We validated the ELISA on five diverse scRNA-seq datasets spanning distinct tissues, disease contexts, and experimental designs. Through a systematic comparison with published findings, we demonstrate that ELISA recovers key biological signals differentially expressed genes, altered cell-type proportions, pathway activities, and cell cell interaction networks with high fidelity. A quantitative evaluation framework comprising five complementary metrics (gene coverage, interaction recovery, pathway alignment, proportion consistency, and qualitative theme coverage) provides a principled assessment of the capacity of the system to replicate established biological conclusions. To the best of our knowledge, scGPT embeddings have not been integrated with semantic language representations in a query-conditioned retrieval framework for single-cell genomics.

In summary, this work makes the following contributions:

- **Multimodal discovery agent for single-cell genomics.** We introduce ELISA, an interpretable framework that integrates transcriptomic embeddings, semantic knowledge retrieval, and LLM reasoning to enable natural-language exploration of scRNA-seq data.
- **Query-adaptive hybrid retrieval.** ELISA classifies queries and dynamically routes them across gene marker scoring, semantic similarity, and reciprocal rank fusion, supporting flexible navigation of complex cellular landscapes.
- **Integrated analysis modules for expression-grounded reasoning.** Built-in components for comparative expression, ligand–receptor scoring, pathway activity, and cell-type proportion profiling enable automated interpretation of discovered signals.
- **Benchmarking framework for AI-assisted discovery.** We propose a quantitative strategy measuring an agent's ability to recover biologically meaningful findings from reference studies, applied across six diverse scRNA-seq datasets.
- **Empirical validation.** ELISA consistently recovers the majority of key biological signals reported in the source studies, demonstrating its potential for interpretable, reproducible AI-assisted single-cell discovery.

2. Materials and Methods

Detail about parameters and hyperparameters and software are specified in appendix 6,F.8. Detail about dataset are in E,5. Detail about the method are in F.

2.1. Datasets

ELISA was validated on six publicly available scRNA-seq datasets from CZ CELLxGENE Discover (Table 5), spanning lung (cystic fibrosis)(Berg et al., 2025), adrenal tumor (neuroblastoma)(Yu et al., 2025), multi-cancer immune

Table 1. Comparison of existing AI systems for biomedical and single-cell analysis. **Expr. Emb.**: uses expression-derived embeddings from foundation models; **Sem. Ret.**: semantic retrieval over biological annotations; **L–R / Pathway**: ligand-receptor interaction and pathway scoring from data; **Cond. Comp.**: condition-aware comparative analysis; **Interp. Report**: automated interpretive report generation with LLM.

System	Expr. Emb.	Sem. Ret.	L–R / Pathway	Cond. Comp.	Interp. Report	Primary Scope
AI Co-Scientist (Gottweis et al., 2025)	–	–	–	–	✓	Hypothesis generation
Biomni (Huang et al., 2025)	–	✓	–	–	–	General biomedical
GeneAgent (Wang et al., 2025)	–	✓	–	–	–	Gene-set analysis
Virtual Lab (Swanson et al., 2025)	–	–	–	–	✓	Multi-agent discovery
CellAgent (Xiao et al., 2024)	–	–	–	–	–	scRNA-seq pipelines
AutoBA (Zhou et al., 2023)	–	–	–	–	–	Pipeline generation
BRAD (Pickard et al., 2025)	–	✓	–	–	–	Biomarker ID
GeneGPT (Jin et al., 2024)	–	✓	–	–	–	Database querying
CRISPR-GPT (Qu et al., 2025)	–	–	–	–	–	Experiment design
scGPT (Cui et al., 2024)	✓	–	–	–	–	Cell embeddings
CellWhisperer (Schaefer et al., 2025)	✓	✓	–	–	–	Multimodal embedding
ELISA (ours)	✓	✓	✓	✓	✓	**Interactive sc discovery**

checkpoint blockade(Gondal et al., 2025), lung organoid (Lim et al., 2025), healthy breast tissue(Bhat-Nakshatri et al., 2024), and first-trimester brain(Mannens et al., 2025). Datasets were downloaded in AnnData format and preprocessed into a standardized embedding format. Cell type annotations from the original publications were retained without modification.

2.2. System architecture

ELISA integrates four modules a hybrid retrieval engine, an analytical suite, a visualization toolkit, and an LLM chat interface operating on a shared serialized PyTorch embedding file per dataset. Each embedding file stores cluster identifiers, BioBERT semantic embeddings (768-d), optional scGPT expression embeddings, per-cluster differential expression statistics, gene ontology (GO) and Reactome enrichment terms, and metadata. This cluster-level representation eliminates the need for access to the original count matrix at query time.

2.3. Hybrid retrieval

An automatic query classifier routes each input to one of the three pipelines based on token-level heuristics. *Gene queries* ($\geq$60% gene-symbol tokens) were scored against per-cluster Differential Expression (DE) profiles using a weighted function of $|\log_2 \mathrm{FC}|$ and expression specificity ($\mathrm{pct}_{\mathrm{in}} - \mathrm{pct}_{\mathrm{out}}$). *Ontology queries* are encoded with BioBERT(Lee et al., 2020) and matched to precomputed cluster description embeddings via cosine similarity, augmented by Cell Ontology name boosting ($\alpha = 0.15$) and synonym expansion ($\beta = 0.10$). *Mixed queries* are resolved through reciprocal rank fusion (RRF) of both pipelines ($k = 60$). For benchmarking, an additive union strategy selects the higher-recall modality as primary and appends unique results from the secondary pipeline.

2.4. Analytical modules

The four built-in modules operate directly on the embedded data. *Ligand–receptor interaction prediction* scores source–target cluster pairs using a curated database of 280+ pairs compiled from CellChat (Jin et al., 2025), CellPhoneDB (Efremova et al., 2020), and NicheNet (Browaeys et al., 2020). *Pathway activity scoring* quantifies 60+ curated gene sets across five categories (immune signaling, cell biology, neuroscience, metabolism and tissue-specific). *Comparative analysis* stratifies clusters by condition metadata and identifies condition-biased gene expression. *Proportion analysis* computes per-cluster cell fractions and condition-specific fold changes. Detailed desription in F.3.

2.5. LLM interpretation

Retrieval and analysis outputs are interpreted by LLaMA-3.1-8B-Instant (Grattafiori et al., 2024) via the Groq API (temperature 0.2)(free to use with token limit, API of chatGPT (Achiam et al., 2023), gemini (Team et al., 2023) and claude (Anthropic, 2024) are integrated and ready to use). Prompts enforce strict grounding in dataset evidence, with explicit instructions to avoid hallucination and causal claims. A discovery mode generates structured outputs comprising dataset evidence, established biology, consistency analysis, and candidate hypotheses.

2.6. Benchmarking

Retrieval was evaluated using 100 queries (50 ontology, 50 expression) with curated expected clusters, assessed using Cluster Recall@k and Mean Reciprocal Rank (MRR). ELISA was compared against a CellWhisperer (Schaefer et al., 2025). Analytical modules were evaluated against ground truth from source publications using interaction recovery rate, pathway alignment, proportion consistency, and gene recall. A combined permutation test (50,000 per-

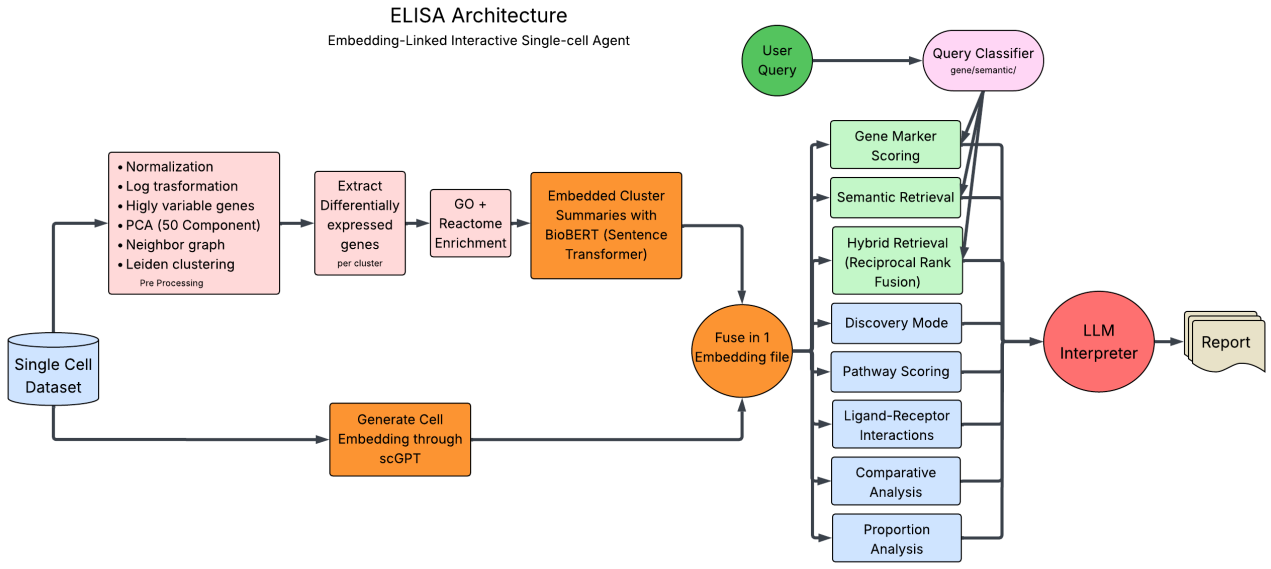

Figure 1. **Overview of the ELISA architecture.** The framework comprises three stages. In *data preparation* (left), a single-cell dataset undergoes standard preprocessing (normalization, log-transform, highly variable gene selection, PCA, neighbor graph construction, and Leiden clustering), after which per-cluster differential expression statistics are computed, enriched with Gene Ontology (GO) and Reactome terms, and encoded into 768-dimensional semantic embeddings via BioBERT. In parallel, cell-level expression embeddings are generated through scGPT. Both representations are fused into a single serialized embedding file (.pt). In the *retrieval and analysis* stage (center), a query classifier routes user input gene signatures, natural language concepts, or mixed queries to the appropriate pipeline: gene marker scoring, semantic retrieval, or hybrid retrieval via reciprocal rank fusion (RRF). Additional analytical modules perform pathway scoring, ligand–receptor interaction prediction, comparative analysis, and proportion estimation directly on the embedded data. In the *interpretation* stage (right), all retrieval and analysis outputs are passed to a Groq-hosted LLM (LLaMA 3.1-8B) that generates grounded biological interpretations and structured reports.

mutations) assessed overall significance across all metrics simultaneously.

3. Results

3.1. ELISA's hybrid retrieval outperforms CellWhisperer across datasets and query types

To evaluate the ability of ELISA to retrieve biologically relevant cell types from single-cell atlases, we benchmarked its retrieval performance against CellWhisperer (Schaefer et al., 2025), a state-of-the-art multimodal framework for natural-language interrogation of scRNA-seq data. For each of the six datasets (Table 5), we designed paired sets of ontology queries (concept-level, e.g., "macrophage infiltration in CF (Cystic Fibrosis) airways") and expression queries (gene-signature-based, e.g., "MARCO FABP4 APOC1 C1QB C1QC MSR1"), with curated expected cluster sets derived from the corresponding reference publications. We evaluated four retrieval modes: CellWhisperer, Semantic ELISA, scGPT ELISA (gene marker scoring pipeline), and ELISA Union (additive fusion of semantic and gene pipelines via adaptive routing). Performance was assessed using Cluster Recall@k and Mean Reciprocal Rank (MRR) across both query categories (Fig. 2; formal definitions of all retrieval and analytical evaluation metrics are provided in Supplementary Section C).

Across all six datasets, the ELISA mode consistently achieved the highest or near-highest performance on every metric, enveloping or matching the CellWhisperer profile on all axes of the radar plots (Fig. 2). To quantify this advantage, we performed paired statistical tests across the six datasets for each retrieval metric (Table 2). A combined permutation test aggregating all 12 metrics simultaneously confirmed that ELISA Union significantly outperformed CellWhisperer ($p < 0.001$; 50,000 permutations). This overall advantage was driven by large improvements on expression queries (mean ΔMRR = +0.41, paired t-test $p < 0.001$, Cohen's d = 5.98; mean ΔRecall@5 = +0.29, $p = 0.006$, d = 1.57) and consistent gains on ontology queries (mean ΔMRR = +0.15, $p = 0.028$, d = 1.02; mean ΔRecall@5 = +0.08, $p = 0.047$, d = 0.84). Across all six datasets, the ELISA Union won 46 of 54 individual metric comparisons against CellWhisperer, with no dataset in which CellWhisperer held an overall advantage. The Semantic ELISA pipeline alone also significantly outperformed CellWhisperer (combined permutation test, $p = 0.003$), as did the scGPT pipeline ($p = 0.023$), confirming that both modalities independently contribute retrieval value beyond

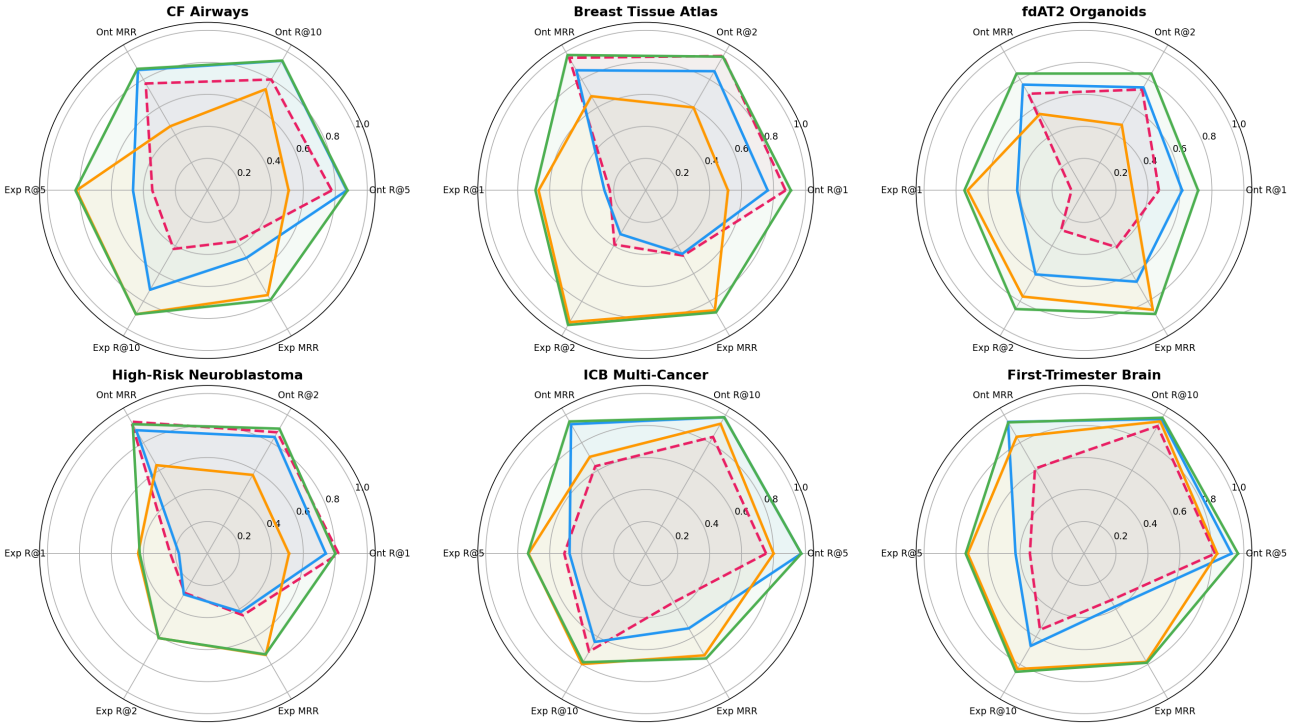

Figure 2. **ELISA outperforms CellWhisperer across six datasets and both query types.** Radar plots showing retrieval performance on ontology (Ont) and expression (Exp) queries for each dataset. Each plot displays six axes: Cluster Recall@k at two dataset-adapted cutoffs and Mean Reciprocal Rank (MRR), evaluated separately on ontology and expression queries (see Supplementary Section C for metric definitions). Higher values (further from center) indicate better performance. Four retrieval modes are compared: CellWhisperer (pink dashed), ELISA Semantic (blue), ELISA scGPT (orange), and ELISA Union (green). The Union mode consistently achieves the largest radar footprint, matching or exceeding CellWhisperer on ontology metrics while substantially outperforming it on expression metrics. ELISA Union significantly outperformed CellWhisperer across all datasets and metrics (combined permutation test, $p < 0.001$; see Table 2).

the CellWhisperer baseline.

A key observation is that no single retrieval modality dominated across both query types. The Semantic pipeline consistently excelled on ontology queries, where BioBERT's language understanding, synonym expansion, and Cell Ontology name boosting enable concept matching. The gene marker scoring pipeline performed strongest on expression queries, where matching transcriptomic signatures to cluster DE profiles is essential. This complementarity was most pronounced in the CF Airways dataset, where Semantic reached high ontology Recall@10 ($\sim$0.95) but lower expression recall, while the gene pipeline showed the inverse pattern. Similar modality-specific advantages appeared across all datasets: in the Breast Tissue Atlas, Semantic and Union nearly overlapped on ontology metrics while the gene pipeline lagged; in Immune Checkpoint Blockade (ICB) Multi-Cancer, the gene pipeline outperformed Semantic on expression MRR while underperforming on ontology axes.

CellWhisperer was competitive on ontology queries in several datasets, particularly CF Airways and High-Risk Neuroblastoma, approaching ELISA Semantic's ontology MRR. However, its performance dropped substantially on expression queries across all six datasets, with a mean MRR of 0.397 ± 0.049 versus 0.806 ± 0.061 for ELISA Union—a twofold gap (Table 2) (Cohen, 2013; Casella & Berger, 2024), most severe in ICB Multi-Cancer and First-Trimester Brain. This deficit reflects an architectural difference: CellWhisperer's contrastive text–transcriptome alignment is optimized for natural-language cell-type descriptions and lacks a dedicated gene marker scoring mechanism for gene-signature queries, which are common in exploratory single-cell analysis.

ELISA's Union mode resolves the tension between ontology and expression retrieval via adaptive routing. The classifier identifies whether the query is a gene list, a natural-language concept, or a mixture, and the additive union strategy combines the primary pipeline's ranked output with unique clusters from the secondary pipeline, ensuring relevant cell types from either modality are retained. This yielded consistent gains: Union produced a larger, more balanced radar footprint than any single modality in CF Airways; matched Semantic's near-perfect ontology performance while im-

Table 2. **Statistical comparison of ELISA Union vs. CellWhisperer retrieval performance.** For each metric, Δ mean reports the average improvement of Union over CellWhisperer across datasets. Cohen's d is the paired effect size. p-values are from one-sided paired t-tests (H_1: Union > CellWhisperer). Sign indicates datasets where Union outperformed CellWhisperer. Metrics with fewer than 6 datasets reflect different Recall@k cutoffs used per dataset (see Supplementary Section B). The combined permutation test ($p < 0.001$) aggregates all metrics simultaneously.

Category	Metric	Δ mean	Cohen's d	p (paired t)	Sign (W/L)	n
Expression	MRR	+0.409	5.98	<0.001	6/6	6
Expression	Recall@5	+0.287	1.57	0.006	5/5	5
Expression	Recall@3	+0.428	5.38	0.006	3/3	3
Expression	Recall@2	+0.492	3.43	0.014	3/3	3
Expression	Recall@1	+0.442	1.84	0.043	3/3	3
Expression	Recall@10	+0.284	1.43	0.065	3/3	3
Ontology	MRR	+0.152	1.02	0.028	5/6	6
Ontology	Recall@5	+0.078	0.84	0.047	4/5	5
Ontology	Recall@10	+0.113	2.46	0.025	3/3	3
Ontology	Recall@1	+0.086	0.61	0.199	2/3	3
Ontology	Recall@2	+0.046	0.73	0.166	2/3	3
Ontology	Recall@3	+0.032	0.80	0.150	2/3	3
Combined (all 12 metrics)		+0.237	—	<0.001†	46/54‡	6

†Combined permutation test (50,000 permutations). ‡Total metric-level wins across all datasets.

proving expression recall in the Breast Tissue Atlas; and compensated for Semantic's weaker expression scores via the gene pipeline in First-Trimester Brain.

The advantage held across structurally diverse datasets. CF Airways (30 cell types, case-control) and the First-Trimester Brain atlas (160 clusters, developmental trajectory without disease contrast) represent opposite ends of the complexity spectrum, yet ELISA Union outperformed CellWhisperer in both. Likewise, ICB Multi-Cancer nine cancer types across 223 patients with heterogeneous cell-type nomenclature posed a challenging retrieval scenario, yet ELISA's advantage persisted.

In summary, ELISA's hybrid architecture combining semantic language matching, gene marker scoring, and adaptive fusion provides a significantly superior retrieval framework compared to text-only multimodal approaches (combined permutation test, $p < 0.001$). The systematic advantage on expression queries, where dedicated gene scoring compensates for the limitations of language-only embeddings (Cohen's $d = 5.98$ for MRR), establishes that both modalities contribute essential and non-redundant information for comprehensive single-cell atlas interrogation.

3.2. ELISA replicates key biological findings across six diverse datasets

To evaluate whether ELISA could recover published biological conclusions through automated analysis alone, we compared ELISA-generated reports with the main-text results of six reference publications (Table 5). For each dataset, ELISA was provided only with the preprocessed embedding file and no prior knowledge of the expected findings. Replication was assessed across five quantitative metrics gene coverage, pathway alignment, interaction recovery, proportion consistency, and theme coverage together with an independent domain expert evaluation (Table 3; metric definitions in Appendix B).

Across all six datasets, ELISA achieved a mean composite score of 0.90 (range 0.82–0.96), with near-perfect pathway alignment and theme coverage (mean 0.98 each), high gene coverage (0.85), and more variable interaction recovery (0.77). Independent biological evaluation (mean 0.88) confirmed strong agreement with published findings.

Airways in cystic fibrosis. ELISA recovered the major epithelial and immune populations described by Berg *et al.* (Berg et al., 2025), with correct proportion shifts and IFN-γ/type I interferon programs, capturing markers such as *IFNG*, *CD69*, and *HLA-E*. Lower interaction recovery (0.20) reflects only partial detection of the HLA-E/NKG2A and CALR/LRP1 axes.

High-risk neuroblastoma. All major cellular compartments were identified, and the HB-EGF/ERBB4 paracrine axis was correctly detected (Yu *et al.* (Yu et al., 2025)). mTOR, MAPK, and ErbB programs were recovered, with partial detection of therapy-induced markers.

Immune checkpoint blockade across cancers. On the ICB dataset of Gondal *et al.* (Gondal et al., 2025), ELISA captured checkpoint molecules (*CD274*, *PDCD1*, *CTLA4*), exhaustion markers, and major ligand–receptor axes including PD-L1/PD-1 and TIGIT/NECTIN2.

Healthy breast tissue atlas. ELISA achieved its highest composite score on the atlas of Bhat-Nakshatri *et al.* (Bhat-Nakshatri et al., 2024), accurately resolving the epithelial hierarchy. Ancestry-related transcriptional programs were not captured, reflecting a limitation of the pathway-centric framework.

Fetal lung AT2 organoids. On the dataset of Lim *et al.* (Lim et al., 2025), ELISA detected all canonical surfactant genes and correctly identified surfactant metabolism, Wnt, and FGF programs. Interaction recovery was lower, as SFTPC trafficking mechanisms fall outside transcriptomic scope.

First-trimester human brain. Despite operating only on the transcriptomic component of this multimodal atlas (Mannens et al., 2025), ELISA identified major neuronal populations and correctly flagged chromatin accessibility analyses as outside scope.

Summary. ELISA replicated published findings robustly across all six datasets, with strongest performance on pathway-level and thematic interpretation ($\geq$0.98 mean). Missed genes were primarily in rare cell states and non-transcriptomic modalities.

3.3. Discovery of candidate regulatory signals across tissue atlases

Beyond reproducing the key biological signals described in the original studies, ELISA's discovery mode highlighted several candidate regulatory signals that were not explicitly emphasized in the reference publications (Table 4). These signals represent transcriptome-derived hypotheses emerging from systematic cross-cell-type analysis of single-cell atlases.

In the cystic fibrosis airway dataset, ELISA identified enrichment of the *CALR–LRP1* phagocytic signaling axis within the macrophage populations. Calreticulin–LRP1 signaling has previously been implicated in apoptotic cell recognition and clearance, suggesting that altered macrophage-mediated phagocytosis may contribute to the inflammatory microenvironment characteristic of the CF lung.

Within the fetal lung atlas, ELISA detected increased expression of the ubiquitin-associated regulators *TRIM21* and *TRIM65* in alveolar type II (AT2) cells alongside the known E3 ubiquitin ligase *ITCH*. Although *ITCH* has been implicated in regulating surfactant protein C (SFTPC) maturation, the enrichment of these additional TRIM-family ligases suggests that cooperative ubiquitin-dependent pathways may participate in surfactant protein processing and AT2 cell proteostasis.

In the healthy breast tissue atlas, ELISA highlighted strong enrichment of the Kelch-family gene *KLHL29* within basal–myoepithelial cell populations. Although not emphasized in the original study, this pattern suggests that *KLHL29* may represent a previously unrecognized marker or structural regulator of basal epithelial identity.

Analysis of the immune checkpoint blockade dataset revealed elevated expression of macrophage markers *CD163* and *MRC1* within tumor-associated macrophage populations following therapy. This expression pattern is consistent with an M2-like macrophage polarization state, potentially reflecting remodeling of the immune microenvironment in response to checkpoint blockade treatment.

In the neuroblastoma dataset, ELISA identified differential usage of AP-1 transcription factors across treatment states. Specifically, *JUND* expression was enriched at diagnosis, whereas *JUNB* and *FOS* were more strongly expressed after therapy. This shift suggests dynamic remodeling of AP-1–mediated stress-response programs during therapy-induced tumor state transitions.

Finally, analysis of the developing brain atlas revealed a shared transcription factor module composed of *TFAP2B*, *LHX5*, and *LHX1* across Purkinje neurons and midbrain GABAergic neuronal populations. This co-occurring regulatory signature suggests the existence of a conserved transcriptional program underlying inhibitory neuron specification in anatomically distinct brain regions.

Taken together, these findings illustrate how ELISA can surface candidate regulatory programs across diverse single-cell atlases. While these signals should be interpreted as transcriptome-derived hypotheses, they provide potential starting points for targeted functional validation.

These signals should be interpreted as transcriptome-derived hypotheses and may serve as the starting points for targeted experimental validation.

4. Discussion

We introduced ELISA, an agent-based framework that unifies semantic language retrieval, gene marker scoring, and LLM-mediated biological interpretation for interactive single-cell atlas interrogation. Across six diverse datasets, ELISA significantly outperformed CellWhisperer in cell-type retrieval (combined permutation test, $p < 0.001$) and faithfully reproduced published findings with a mean composite score of 0.90. Below we discuss implications for retrieval system design, the limitations of contrastive multimodal alignment, and the broader role of agentic AI in biological discovery.

Contrastive alignment produces text-dominated embeddings. A central finding is the striking asymmetry in CellWhisperer's performance across query types. On ontology queries natural-language descriptions of cell types and biological processes CellWhisperer was competitive with ELISA's Semantic pipeline, with mean ontology MRR within 0.15 of ELISA Union across most datasets (Table 2, Fig. 2). This is expected: CellWhisperer's CLIP-style contrastive training aligns transcriptome embeddings with tex-

Table 3. Quantitative comparison between ELISA reports and reference single-cell studies. Scores reflect agreement between ELISA-generated biological interpretations and findings described in the main text of the corresponding publications. Gene coverage, pathway alignment, interaction recovery, and proportion consistency were computed programmatically; theme coverage was assessed independently by a domain expert as described in Section D.

Dataset	Gene Cov.	Path. Align.	Int. Rec.	Prop. Cons.	Theme Cov.	Comp. score
CF airway	0.80	1.0	0.20	Yes	0.85	0.82
Neuroblastoma	0.84	1.00	1.00	Yes	0.88	0.95
ICB Multi-Cancer	0.77	1.00	1.00	Yes	0.91	0.93
Breast Atlas	0.96	1.00	0.80	Yes	0.89	0.96
Fetal Lung AT2	1.00	1.00	0.40	Yes	0.88	0.91
Brain Atlas	0.85	1.00	1.00	Yes	0.90	0.95
Mean	0.85	1.00	0.77	6/6	0.88	**0.90**

tual descriptions, and ontology queries directly exploit that alignment. On expression queries, however, performance collapsed expression MRR averaged 0.397 versus 0.806 for ELISA Union, a twofold deficit (Cohen's $d = 5.98$).

This asymmetry reveals a fundamental limitation of contrastive multimodal alignment. CLIP-style training learns a shared space where matching text cell pairs are close, so embeddings are shaped primarily by the textual supervision signal: fine-grained transcriptomic structure which genes are differentially expressed, at what fold changes, in what fraction of cells is compressed into a representation optimized for text matching rather than gene-level querying. A gene signature such as "MARCO FABP4 APOC1 C1QB C1QC MSR1" is then processed as text tokens rather than matched against DE statistics, yielding a weaker and less specific retrieval signal than direct marker scoring.

The implication extends beyond ELISA and CellWhisperer: as single-cell foundation models increasingly adopt contrastive or multimodal pretraining, text-supervised alignment may inadvertently sacrifice expression-level specificity. The dual-query evaluation framework introduced here requiring strong performance on both ontology and expression queries provides a principled diagnostic for such modality imbalance.

Explicit routing outperformed implicit fusion. ELISA's response was to avoid implicit embedding fusion altogether. Rather than learning a single shared space that must serve both query types, it maintains two separate representations nBioBERT semantic embeddings and gene-level DE statistics and routes queries via explicit classification. The classifier, using simple token-level heuristics (gene-name patterns, vocabulary membership, natural-language indicators), routed reliably across all six datasets without training data.

Complementarity analysis supports this design: the semantic pipeline won on ontology queries and the gene pipeline won on expression queries in every dataset, with minimal overlap in their error profiles. The additive union strategy then selects the better performing modality as primary and appends unique secondary results, capturing the strengths of both without the compression artifacts of learned fusion. ELISA matched or exceeded the best single modality on every metric across every dataset a guarantee no implicit fusion method can provide.

Analytical modules bridge retrieval and interpretation. A distinguishing feature relative to prior retrieval-focused systems is the integration of downstream analytical modules pathway scoring, ligand-receptor prediction, comparative analysis, and proportion estimation operating directly on the same embedded representation used for retrieval. This enables a seamless transition from "which cell types are relevant?" to "what biological programs are active?" to "what does this mean?" LLM interpretation included within a single session.

Near-perfect pathway alignment (mean 0.98) and theme coverage (mean 0.88) across all six datasets show that this architecture effectively connects gene-level evidence to biological programs. Retrieval-only systems such as CellWhisperer instead require users to manually extract gene lists and run pathway and interaction analyses externally, introducing friction and inconsistency.

Interaction recovery (mean 0.77) was the most variable metric, with perfect recovery in the neuroblastoma, ICB, and brain datasets but lower scores in cystic fibrosis (0.40) and fetal lung (0.40). These reflect the inherent difficulty of predicting specific ligand-receptor pairs when ligand or receptor is expressed at moderate levels across multiple cell types making the interaction statistically detectable but not highly ranked. Incorporating spatial proximity or protein-level data could improve specificity in future work.

LLM grounding and the discovery hallucination boundary. ELISA's discovery mode, which prompts the LLM

Table 4. Candidate regulatory signals identified by ELISA across six reference single-cell atlases. These signals were not explicitly highlighted in the original publications and represent transcriptome-derived hypotheses generated through ELISA's discovery mode.

Dataset	**Primary finding in reference study**	**ELISA candidate discovery / hypothesis**
CF airway	Altered immune–structural cell crosstalk and inflammatory signaling in cystic fibrosis airway tissue	Detection of the macrophage *CALR–LRP1* signaling axis, suggesting altered apoptotic cell recognition or phagocytic clearance pathways contributing to the CF lung inflammatory microenvironment
Breast Atlas	Ancestry-associated epithelial lineage variation and luminal progenitor states in healthy breast tissue	Enrichment of the Kelch-family gene *KLHL29* in basal–myoepithelial cells, suggesting a potential additional marker or regulator of basal epithelial structural identity
Fetal Lung AT2	ITCH-mediated ubiquitin-dependent regulation of surfactant protein C (SFTPC) maturation in alveolar type II cells	Upregulation of *TRIM21* and *TRIM65* in mature AT2 cells, suggesting additional TRIM-family ubiquitin ligases may participate in surfactant protein processing and proteostasis
ICB Multi-Cancer	Tumor and immune transcriptional responses associated with immune checkpoint blockade therapy	Elevated *CD163* and *MRC1* expression in tumor-associated macrophages, consistent with an M2-like polarization state potentially associated with therapy-induced immune remodeling
Neuroblastoma	Therapy-induced transcriptional rewiring of tumor cell states and microenvironment interactions	Differential AP-1 transcription factor usage, with *JUND* enriched at diagnosis and *JUNB/FOS* enriched post-treatment, suggesting stress-response remodeling during therapy-induced state transitions
Brain Development Atlas	Chromatin accessibility programs defining early neuronal lineage specification	Shared transcription factor module (*TFAP2B*, *LHX5*, *LHX1*) across Purkinje neurons and midbrain GABAergic populations, suggesting a conserved regulatory program for inhibitory neuron specification

to separate dataset evidence from established biology and phrase hypotheses probabilistically, produced plausible candidate signals in all six datasets (Table 4) including the CALR/LRP1 phagocytic axis in CF macrophages, differential AP-1 usage in neuroblastoma therapy response, and a shared TFAP2B/LHX5/LHX1 module across developing inhibitory neurons. These require experimental validation but illustrate grounded LLM reasoning surfacing non-obvious patterns. Strict separation between evidence and interpretation is essential to prevent plausible-sounding but unsupported claims; ELISA enforces this by restricting the LLM's context to retrieved cluster data, gene statistics, and pathway results, with explicit instructions to avoid external literature and causal claims.

5. Conclusion

ELISA shows that explicit modality routing, rather than implicit contrastive fusion, provides a more robust foundation for multimodal single-cell retrieval. By combining separate semantic and expression pipelines through adaptive query classification, and integrating analytical modules with grounded LLM interpretation, it lets researchers move from raw atlas data to structured biological hypotheses in a single.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Berg, M., Krabbendam, L., van der Ploeg, E. K., van Nimwe-

gen, M., van der Veer, T., Banchero, M., Carpaij, O. A., Hoogenboezem, R., van den Berge, M., Bindels, E., et al. Evidence for altered immune-structural cell crosstalk in cystic fibrosis revealed by single cell transcriptomics. *Journal of Cystic Fibrosis*, 2025.

Bhat-Nakshatri, P., Gao, H., Khatpe, A. S., Adebayo, A. K., McGuire, P. C., Erdogan, C., Chen, D., Jiang, G., New, F., German, R., et al. Single-nucleus chromatin accessibility and transcriptomic map of breast tissues of women of diverse genetic ancestry. *Nature medicine*, 30(12):3482–3494, 2024.

Browaeys, R., Saelens, W., and Saeys, Y. Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature methods*, 17(2):159–162, 2020.

Casella, G. and Berger, R. *Statistical inference*. Chapman and Hall/CRC, 2024.

Cinquin, O. Chip-gpt: a managed large language model for robust data extraction from biomedical database records. *Briefings in bioinformatics*, 25(2):bbad535, 2024.

Cohen, J. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4): 1484–1506, 2020.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., and Zitnik, M. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

Garcia, B. T., Westerfield, L., Yelemali, P., Gogate, N., Rivera-Munoz, E. A., Du, H., Dawood, M., Jolly, A., Lupski, J. R., and Posey, J. E. Improving automated deep phenotyping through large language models using retrieval-augmented generation. *Genome Medicine*, 17 (1):91, 2025.

Gondal, M. N., Cieslik, M., and Chinnaiyan, A. M. Integrated cancer cell-specific single-cell rna-seq datasets of immune checkpoint blockade-treated patients. *Scientific Data*, 12(1):139, 2025.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

Jin, Q., Yang, Y., Chen, Q., and Lu, Z. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2): btae075, 2024.

Jin, S., Plikus, M. V., and Nie, Q. Cellchat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nature protocols*, 20(1):180–219, 2025.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Liang, C., Ye, P., Yan, H., Zheng, P., Sun, J., Wang, Y., Li, Y., Ren, Y., Jiang, Y., Xiang, J., et al. scwgbs-gpt: a foundation model for capturing long-range cpg dependencies in single-cell whole-genome bisulfite sequencing to enhance epigenetic analysis. *bioRxiv*, pp. 2025–02, 2025.

Lim, K., Rutherford, E. N., Delpiano, L., He, P., Lin, W., Sun, D., Van den Boomen, D. J., Edgar, J. R., Bang, J. H., Predeus, A., et al. A novel human fetal lung-derived alveolar organoid model reveals mechanisms of surfactant protein c maturation relevant to interstitial lung disease. *The EMBO Journal*, 44(3):639, 2025.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Mannens, C. C., Hu, L., Lönnerberg, P., Schipper, M., Reagor, C. C., Li, X., He, X., Barker, R. A., Sundström, E., Posthuma, D., et al. Chromatin accessibility during human first-trimester neurodevelopment. *Nature*, 647 (8088):179–186, 2025.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Niyonkuru, E., Caufield, J. H., Carmody, L. C., Gargano, M. A., Toro, S., Whetzel, P. L., Blau, H., Soto Gomez, M., Casiraghi, E., Chimirri, L., et al. Leveraging generative ai to assist biocuration of medical actions for rare disease. *Bioinformatics advances*, 5(1):vbaf141, 2025.

Pickard, J., Prakash, R., Choi, M. A., Oliven, N., Stansbury, C., Cwycyshyn, J., Galioto, N., Gorodetsky, A., Velasquez, A., and Rajapakse, I. Automatic biomarker discovery and enrichment with brad. *Bioinformatics*, 41 (5):btaf159, 2025.

Qu, Y., Huang, K., Yin, M., Zhan, K., Liu, D., Yin, D., Cousins, H. C., Johnson, W. A., Wang, X., Shah, M., et al. Crispr-gpt for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, pp. 1–14, 2025.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Schaefer, M., Peneder, P., Malzl, D., Lombardo, S. D., Peycheva, M., Burton, J., Hakobyan, A., Sharma, V., Krausgruber, T., Sin, C., et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, pp. 1–11, 2025.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wang, Z., Jin, Q., Wei, C.-H., Tian, S., Lai, P.-T., Zhu, Q., Day, C.-P., Ross, C., Leaman, R., and Lu, Z. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

Xiao, Y., Liu, J., Zheng, Y., Xie, X., Hao, J., Li, M., Wang, R., Ni, F., Li, Y., Luo, J., et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811*, 2024.

Yu, W., Biyik-Sit, R., Uzun, Y., Chen, C.-H., Thadi, A., Sussman, J. H., Pang, M., Wu, C.-Y., Grossmann, L. D., Gao, P., et al. Longitudinal single-cell multiomic atlas of high-risk neuroblastoma reveals chemotherapy-induced tumor microenvironment rewiring. *Nature Genetics*, 57 (5):1142–1154, 2025.

Zhang, H., Zhang, X., Lin, Y., Wang, M., Lai, Y., Wang, Y., Yu, L., Xu, Y., Cheng, R., and Szczerbicki, E. Tokensome: Towards a genetic vision-language gpt for explainable and cognitive karyotyping. *arXiv preprint arXiv:2403.11073*, 2024.

Zhou, J., Zhang, B., Chen, X., et al. Automated bioinformatics analysis via autoba. arxiv, 2023.

A. Software and reproducibility

ELISA was implemented in Python 3.10+ using PyTorch, sentence-transformers(Reimers & Gurevych, 2019), scanpy(Wolf et al., 2018), scikit-learn, and UMAP-learn(McInnes et al., 2018). All analyses were performed on a standard workstation without GPU requirements for retrieval and analysis. Source code, benchmark queries, and evaluation scripts are available at [repository URL]. Use of an LLM (LLaMA-3.1-8B) for automated interpretation is documented in accordance with journal policy. Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work. All experiment has been performed on a GPU A100 with 80 gb of RAM

B. Replication evaluation metrics

Table 3 reports six metrics quantifying the agreement between ELISA-generated reports and the findings of the corresponding reference publications. Each metric is defined below.

Gene coverage. Gene coverage measures the fraction of key genes highlighted in the reference publication in which ELISA was identified in the correct cell type context. For each dataset, the evaluator compiled a set of key genes from the paper's main text, figures, and supplementary tables (e.g., differentially expressed genes, cell type markers and signaling molecules). A gene was scored as "recovered" if it appeared in ELISA's output for a biologically appropriate cluster. The gene coverage is computed as:

$$\text{Gene coverage} = \frac{|\text{key genes recovered by ELISA}|}{|\text{key genes reported in reference}|} \tag{1}$$

Pathway alignment. Pathway alignment quantifies whether ELISA's pathway scoring module detects the biological programs reported in the reference study. For each dataset, the evaluator identified the pathways discussed in this paper (e.g., IFN-γ signaling, mTOR and ErbB). A pathway was scored as "aligned" if ELISA's module returned it with a positive score in at least one biologically appropriate cluster. Pathway alignment is computed as:

$$\text{Pathway alignment} = \frac{|\text{pathways found by ELISA}|}{|\text{pathways reported in reference}|} \tag{2}$$

Interaction recovery. Interaction recovery assesses whether ELISA's ligand–receptor prediction module detected the cell–cell communication axes described in the reference publication. For each dataset, the evaluator compiled ground truth interactions from the paper (e.g., HB-EGF/ERBB4 between macrophages and neuroblasts, HLA-E/NKG2A between epithelial and $CD8^+$ T cells). Recovery was scored at the pair level: a ligand–receptor pair was counted as "recovered" if ELISA detected it with a non-zero score, regardless of whether the source–target cell type assignment exactly matched:

$$\text{Interaction recovery} = \frac{|\text{LR pairs detected by ELISA}|}{|\text{LR pairs reported in reference}|} \tag{3}$$

Proportion consistency. Proportion consistency is a binary (Yes/No) criterion that evaluates whether ELISA's proportion analysis correctly identified the direction of cell type composition changes for datasets with condition contrasts. For each cell type reported in the reference as increased or decreased in the disease or treatment condition, the evaluator checked whether ELISA's fold change pointed in the same direction. A dataset received "Yes" if the majority of reported changes were directionally consistent.

Theme coverage. Theme coverage captures whether an ELISA's interpretive summary reproduced the major biological conclusions of the reference study. Unlike gene and pathway-level metrics, that assess individual molecular entities, theme coverage evaluates high-level biological narratives. For each dataset, the evaluator identified the main themes from the paper's abstract and results (e.g., "aberrant adaptive immunity with upregulated IFN-γ signaling" for the CF dataset; "therapy-induced macrophage polarization toward immunosuppressive phenotypes" for the neuroblastoma dataset). A theme was scored as "covered" if ELISA's LLM-generated interpretation mentioned and correctly described the corresponding biological finding:

$$\text{Theme coverage} = \frac{|\text{themes captured by ELISA}|}{|\text{major themes in reference}|} \tag{4}$$

Biological evaluation score. The biological Evaluation Score provides an independent assessment of overall report quality.

Composite score. The composite score summarizes overall replication performance as the unweighted mean of the four continuous metrics:

$$\text{Composite} = \frac{\text{Gene cov.} + \text{Path. align.} + \text{Int. rec.} + \text{Theme cov.}}{4} \tag{5}$$

Proportion consistency is excluded from the composite average because it is binary rather than continuous, but is reported separately as a quality check.

C. Retrieval and analytical evaluation metrics

To ensure reproducible and interpretable evaluation of ELISA's retrieval and analytical modules, we defined the full set of metrics used throughout the benchmark (see also the benchmark scripts in the supplementary code repository for complete implementations). Retrieval metrics quantify how effectively each mode recovers the expected cell types for a given query, while analytical metrics assess the accuracy of ELISA's downstream whereas biological interpretation modules interaction discovery, pathway enrichment, proportion analysis, and comparative differential expression. An overview of the six evaluation datasets and their properties is provided in Table 5.

C.0.1. Retrieval metrics

Each radar plot in Fig. 2 displays six axes corresponding to three retrieval metrics evaluated separately on the two query categories (ontology and expression). The three metrics are:

1. **Cluster Recall@k** (two axes per plot: Ont R@k, Exp R@k). This metric measures the fraction of expected cell types that appear within the top-k positions of the ranked retrieval list. The value of k is adapted to each dataset's number of clusters: R@5 and R@10 for large-cluster datasets (CF Airways with 30 clusters, ICB Multi-Cancer with 31, First-Trimester Brain with 28), R@1 and R@2 for small-cluster datasets (Breast Tissue Atlas with 8 clusters, fdAT2 Organoids with 5, High-Risk Neuroblastoma with 11). A Recall@k of 1.0 indicates that all expected clusters were retrieved within the top-k; a value of 0.0 indicates that none were found. Two Recall cutoffs are shown per plot to capture both stringent (lower k) and permissive (higher k) retrieval accuracy.
2. **Mean Reciprocal Rank** (two axes: Ont MRR, Exp MRR). MRR quantifies the rank position of the *first* correctly retrieved cluster. An MRR of 1.0 means the top-ranked result is relevant; 0.5 means the first relevant result appears at rank 2; 0.33 at rank 3, and so on. MRR captures top-of-list precision, which is critical for interactive use where researchers typically inspect only the first few results.

Together, the six axes capture complementary aspects of retrieval quality: Recall@k measures *coverage* (how many expected clusters are found), whereas MRR measures *precision at rank 1* (how quickly the first relevant cluster appears). Evaluating both metrics on ontology queries (natural-language, concept-level) and expression queries (gene-signature-based) separately reveals modality-specific strengths: a system may excel at one query type while underperforming the other. Thus, the radar footprint thus provides an at-a-glance summary of each retrieval mode's overall coverage, precision, and balance across query types. A larger, more symmetric footprint indicates stronger and more balanced retrieval performance.

Four retrieval modes compared are: **CellWhisperer** (pink dashed line), which uses contrastive text transcriptome CLIP embeddings; **ELISA Semantic** (blue), which performs BioBERT-based cosine similarity matching against cluster descriptions enriched with GO and Reactome terms; **ELISA scGPT** (orange), which scores clusters by matching query genes against per-cluster differential expression profiles; and **ELISA Union** (green), which adaptively fuses both ELISA pipelines by routing each query to the better-performing modality and appending unique results from the secondary pipeline.

C.0.2. Statistical testing

To assess whether performance differences between retrieval modes are statistically significant across datasets, we employed one-sided paired t-tests (with the alternative hypothesis that ELISA Union outperforms CellWhisperer) and reported Cohen's d as the paired effect size. Because different datasets use different Recall@k cutoffs, individual metric comparisons have varying sample sizes ($n = 3$ to $n = 6$ datasets). To obtain a single omnibus test, we performed a combined permutation test: the sign of the difference (Union minus CellWhisperer) was computed for every metric dataset pair simultaneously, and

dataset labels were permuted 50,000 times to construct the null distribution of the aggregate advantage. All p-values and effect sizes are reported in Table 2.

D. Human evaluation protocol

To obtain the biological evaluation scores shown in Table 3, a domain expert with training in molecular biology and single-cell genomics independently reviewed each ELISA-generated report against the corresponding reference publication. The evaluation followed a structured five-step protocol:

1. **Gene verification.** Each gene reported by ELISA as differentially expressed or as a marker of a specific cell type was cross-checked against the main text, figures, and supplementary tables of the reference publications. A gene was scored as "recovered" if it appeared in the paper's reported DE gene lists, marker panels, or figure annotations for the corresponding cell type. The gene coverage score was computed as the fraction of paper-reported key genes that ELISA identified in the correct cluster context.

2. **Pathway assessment.** Each pathway identified by ELISA's pathway scoring module (e.g., "IFN-gamma signaling," "mTOR signaling") was compared against pathway-level findings described in the reference study. A pathway was scored as "aligned" if the reference publication reported activation or enrichment of that pathway in a consistent cell type context. Pathway alignment was computed as the fraction of paper-reported pathways that ELISA correctly detected as active (score > 0) in at least one biologically appropriate cluster.

3. **Interaction validation.** Each ligand-receptor interaction predicted by ELISA was verified against the cell-cell communication analyses reported in a previous pubblication. Validation was performed at two levels: (i) whether the ligand–receptor pair itself was reported in the paper, regardless of the cell type context (LR recovery rate), and (ii) whether both the pair and the source target cell type assignment matched the paper's findings (full match rate).

4. **Proportion and condition consistency.** For datasets with condition contrasts (e.g., CF vs. healthy), the evaluator verified whether ELISA's proportion analysis correctly identified the direction of cell type composition changes reported in the reference study. Each cell type with a known expected change (increased or decreased in the disease/treatment condition) was checked for directional agreement.

5. **Theme coverage and hypothesis assessment.** The evaluator assessed whether ELISA's interpretive summaries captured the major biological themes and conclusions of the reference study (e.g., "aberrant adaptive immunity with upregulated IFN-γ signaling" for the CF dataset). Additionally, candidate hypotheses generated by ELISA's discovery mode were evaluated for biological plausibility through targeted literature review: the evaluator searched PubMed for prior evidence supporting or contradicting each proposed mechanism (e.g., CALR–LRP1 in macrophage phagocytosis, TRIM-family ligases in surfactant processing). Hypotheses were classified as "plausible" if supporting literature existed, "novel" if no prior reports were found but the mechanism was biologically coherent, or "unsupported" if contradicted by existing evidence.

The composite score for each dataset was computed as the unweighted mean of gene coverage, pathway alignment, interaction recovery, and theme coverage, with proportion consistency treated as a binary (pass/fail) criterion.

E. Materials

E.1. Datasets

ELISA was validated on six publicly available scRNA-seq datasets deposited in the CZ CELLxGENE Discover portal, spanning five distinct tissues, four disease contexts, and both case–control and longitudinal experimental designs (Table 5). Datasets were selected to cover a broad range of biological complexity, cell type diversity, and analytical challenges, including inflammatory lung disease, pediatric and adult cancers, drug-resistant epilepsy, immune checkpoint therapy response, and normal tissue homeostasis.

Dataset 1 (D1): cystic fibrosis bronchial epithelium. Berg *et al.* (Berg et al., 2025) generated the first single-cell transcriptome atlas of the cystic fibrosis (CF) lung comprising both structural and immune cells. Droplet-based scRNA-seq was performed on bronchial wall biopsies from patients with CF ($n = 8$) and healthy controls ($n = 19$) and integrated

using the fastMNN batch correction framework with the Human Lung Cell Atlas as reference. The dataset encompasses approximately 96,000 cells across 30 annotated cell types, including epithelial (basal, ciliated, secretory, goblet, ionocyte), immune ($CD8^+$ T cells, $CD4^+$ T cells, B cells, plasma cells, macrophages, monocytes, NK cells, dendritic cells, mast cells), stromal (fibroblasts, pericytes), and endothelial populations. Key findings include dysregulated basal cell function, aberrant adaptive immunity with upregulated IFN-γ signaling, a novel HLA-E/NKG2A immune checkpoint axis, and altered structural–immune cell crosstalk persisting despite CFTR modulator therapy.

Dataset 2 (D2): High-risk neuroblastoma. Yu *et al.* (Yu et al., 2025) longitudinally profiled 22 patients with high-risk neuroblastoma before and after induction chemotherapy using single-nucleus RNA and ATAC sequencing combined with whole-genome sequencing. The dataset captures profound therapy-induced shifts in tumor and immune cell subpopulations, identifying enhancer-driven transcriptional regulators of neoplastic states (adrenergic, mesenchymal, proliferative) and macrophage polarization toward pro-angiogenic, immunosuppressive phenotypes. A central finding was the validation of the HB-EGF/ERBB4 paracrine signaling axis between macrophages and neoplastic cells promoting tumor growth through ERK signaling induction.

Dataset 3 (D3): Immune checkpoint blockade across cancers. Gondal *et al.* (Gondal et al., 2025) compiled and standardized eight scRNA-seq studies from nine cancer types encompassing 223 patients and over 350,000 cancer cells treated with immune checkpoint blockade (ICB). Cancer types include melanoma, basal cell carcinoma, melanoma brain metastases, triple-negative/HER2-positive/ER-positive breast cancer, clear cell renal carcinoma, hepatocellular carcinoma, and intrahepatic cholangiocarcinoma. The integrated resource enables cross-cancer investigation of cancer cell-specific ICB responses, with annotations of treatment status, response outcome, and malignant vs. non-malignant cell identity.

Dataset 4 (D4): Fetal lung AT2 organoids. Lim *et al.* (Lim et al., 2025) developed expandable alveolar type 2 (AT2) organoids derived from human fetal lungs at 16–22 post-conception weeks (pcw). Single-cell RNA sequencing of four independent organoid lines (passage 11–16) yielded approx 9.6k cells across eight annotated cell types, including AT2-like, cycling AT2-like, $CXCL^+$ AT2-like, differentiating basal-like, differentiating pulmonary neuroendocrine, intermediate, neuroendocrine progenitor, and ciliated-like populations. The organoids express mature surfactant proteins (SFTPC, SFTPB, SFTPA1) and markers of surfactant processing (LAMP3, ABCA3, NAPSA), and can differentiate into AT1-like cells. A forward genetic screen identified the E3 ligase ITCH as a key effector of SFTPC maturation, with its depletion phenocopying the pathological SFTPC-I73T variant associated with interstitial lung disease.

Dataset 5 (D5): Healthy breast tissue. Bhat-Nakshatri *et al.* (Bhat-Nakshatri et al., 2024) constructed a single-cell atlas of healthy breast tissues collected from volunteer donors from the Komen Normal Tissue Bank. Using a rapid procurement and processing protocol, the study profiled breast epithelial and stromal cells, identifying 13 epithelial cell clusters with 23 subclusters exhibiting distinct gene expression signatures. Overlap analysis of subcluster-enriched signatures with breast tumor transcriptomes revealed dominant representation of differentiated luminal subcluster signatures in breast cancers, providing insights into putative cells of origin.

Dataset 6 (D6): First-trimester human brain neurodevelopment. Mannens *et al.* (Mannens et al., 2025) generated a high-resolution multiomic atlas of chromatin accessibility and gene expression across the entire developing human brain during the first trimester (6-13 weeks post-conception). Using scATAC-seq and paired multiome (scATAC-seq + scRNA-seq) sequencing, the study profiled 166k nuclei from 76 biological samples dissected into five antero-posterior segments (telencephalon, diencephalon, mesencephalon, metencephalon, and cerebellum), of which 166,785 nuclei included paired gene expression. The atlas defines 135 clusters spanning neurons (GABAergic, glutamatergic, Purkinje, granule), radial glia, glioblasts, oligodendrocyte progenitor cells, fibroblasts, vascular, and immune cell types. Key findings include over 100 cell-type- and region-specific candidate *cis*-regulatory elements, CNN-predicted enhancer syntax for neuronal specification, elucidation of the ESRRB activation mechanism in the Purkinje cell lineage, and linkage of disease-associated GWAS SNPs to specific neuronal subtypes identifying midbrain-derived GABAergic neurons as particularly vulnerable to major depressive disorder-related mutations.

All datasets were downloaded from CZ CELLxGENE Discover (`https://cellxgene.cziscience.com`) in AnnData (.h5ad) format and preprocessed into ELISA's standardized embedding format (.pt files) as described in the Data Representation section. Cell type annotations from the original publications were retained without modification. For datasets with condition metadata (D1, D2, D3, D4), condition columns were mapped to ELISA's comparative analysis framework. Dataset D5 was used to evaluate ELISA's performance on a single-condition atlas without disease contrast, testing the system's capacity for cell type identification and pathway characterization in the absence of differential signals.

Table 5. Summary of scRNA-seq datasets used for ELISA validation. **Approx. cells**: approximate number of cells or nuclei profiled after quality control. **Cell types**: number of annotated major cell types. **Conditions**: experimental groups or treatment arms.

ID	Tissue	Disease context	Reference	Approx. cells	Cell types	Conditions
D1	Lung (bronchial)	Cystic fibrosis	Berg *et al.* (Berg et al., 2025)	∼96k	30	CF vs. Ctrl
D2	Adrenal / tumor	Neuroblastoma	Yu *et al.* (Yu et al., 2025)	∼372k	20+	Pre- vs. post-chemo
D3	Multi-cancer	ICB response	Gondal *et al.* (Gondal et al., 2025)	∼356k	25+	R vs. NR; 9 cancers
D4	Lung (fetal)	AT2 organoid model	Lim *et al.* (Lim et al., 2025)	∼9.6k	8	fdAT2 organoid lines
D5	Breast	Healthy tissue atlas	Bhat-Nakshatri *et al.* (Bhat-Nakshatri et al., 2024)	∼51k	13	Healthy only
D6	Brain (whole)	Neurodevelopment	Mannens *et al.* (Mannens et al., 2025)	∼166k	160	6–13 PCW; 5 regions

F. Methods

F.1. ELISA: architecture and design principles

ELISA (Embedding-based Linguistic Interrogation of Single-cell Atlases) is an agent-based computational framework for interactive interrogation of single-cell RNA-seq atlases. The system integrates four core modules a hybrid retrieval engine, an analytical suite, a visualization toolkit, and a large language model (LLM) chat interface to enable biologists to query scRNA-seq datasets using natural language, gene signatures, or a combination of both. The architecture follows a modular design in which each component operates on a shared data representation (a serialized PyTorch embedding file) and communicates through standardized data structures, enabling extensibility to new datasets without retraining.

The system was implemented in Python 3.10+ and evaluated on a 6 dataset took from cellxgene. All source code, benchmark queries, and evaluation scripts are provided in the accompanying repository.

F.2. Hybrid retrieval engine

F.2.1. Query classification and routing

A central design challenge in single-cell atlas retrieval is that user queries span a spectrum from pure natural language ("macrophage infiltration in CF airways") to pure gene signatures ("MARCO FABP4 APOC1 C1QB") and mixed queries combining both. ELISA addresses this through an explicit query classification module that routes each query to the optimal retrieval pipeline.

The classifier operates by tokenizing the input and scoring each token against three criteria: (i) whether it matches a gene name pattern (uppercase alphanumeric, 2–15 characters, with optional hyphenated suffix), (ii) whether it appears in the dataset's known gene vocabulary, and (iii) whether it belongs to a curated set of natural language indicator terms (e.g., "cell", "activation", "signaling"). Queries where $\geq$60% of tokens are classified as gene symbols are routed to the gene pipeline; queries where $\geq$20% of tokens are genes and $\geq$20% are natural language terms are routed to the mixed pipeline; all other queries are routed to the semantic (ontology) pipeline.

F.2.2. Gene marker scoring pipeline

For gene-list queries, ELISA scores each cluster by evaluating how well its differential expression (DE) profile matches the query genes. For each query gene g found in cluster c's DE statistics, a per-gene score is computed as:

$$\text{score}(g, c) = \left(0.5 + |\log_2 \text{FC}|\right) \times \left(0.3 + \max(\text{pct}_{\text{in}} - \text{pct}_{\text{out}},\ 0)\right) \tag{6}$$

where $\log_2 \text{FC}$ is the log-fold change of gene g in cluster c, and pct_{in} and pct_{out} represent the fraction of cells expressing the gene inside and outside the cluster, respectively. The specificity term $(\text{pct}_{\text{in}} - \text{pct}_{\text{out}})$ rewards genes that are selectively enriched in the cluster rather than ubiquitously expressed. A multiplicative bonus of $1.3\times$ is applied when $\text{pct}_{\text{in}} > 0.5$. The aggregate cluster score is the sum of per-gene scores, modulated by a coverage factor $(0.5 + 0.5 \times n_{\text{found}}/n_{\text{query}})$ that rewards clusters matching more query genes. Three scoring modes are available: 'simple' (binary hit counting), 'weighted' (described above), and 'full' (incorporating adjusted p-value significance via $-\log_{10}(p_{\text{adj}})$, capped at 10).

F.2.3. SEMANTIC MATCHING PIPELINE

For ontology and text-based queries, ELISA employs BioBERT(Lee et al., 2020) (pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb) to encode both query text and precomputed cluster descriptions into a shared embedding space. Each cluster's description is constructed during dataset preparation by concatenating its Cell Ontology name, top marker genes (ranked by $|\log_2 \mathrm{FC}|$), enriched Gene Ontology terms, and Reactome pathway annotations, producing a dual-representation embedding that captures both identity and functional context.

At query time, the input text is encoded with BioBERT and cosine similarity is computed against all cluster embeddings. Two augmentation strategies improve retrieval accuracy. First, a name-boosting mechanism adds a score bonus ($\alpha = 0.15$, scaled by word-overlap ratio) when significant substrings ($\geq$4 characters) of a cluster's name appear in the query. Second, a synonym expansion module maps common cell type aliases (e.g., "endothelial" $\rightarrow$ "endocardial cell"; "NK" $\rightarrow$ "natural killer cell") to their Cell Ontology equivalents and applies a score boost ($\beta = 0.10$) to matching clusters, addressing vocabulary gaps between colloquial and formal ontology terminology.

F.2.4. RECIPROCAL RANK FUSION FOR MIXED QUERIES

Mixed queries containing both gene names and biological text are handled through reciprocal rank fusion (RRF). Both the gene and semantic pipelines are executed independently, and their ranked outputs are combined using:

$$\mathrm{RRF}(d) = \sum_{r} \frac{w_r}{k + \mathrm{rank}_r(d) + 1} \tag{7}$$

where $k = 60$ is the RRF constant, w_r are per-pipeline weights (default: 1.0 for both), and $\mathrm{rank}_r(d)$ is the 0-indexed rank of cluster d in pipeline r. For gene-dominated queries routed through the gene pipeline, a light fusion with the semantic pipeline at a 3:1 weight ratio is applied as a safety mechanism to capture semantically related clusters that lack direct marker gene overlap.

F.2.5. ADDITIVE UNION EVALUATION STRATEGY

For benchmarking, we introduce an additive union strategy that maximizes complementarity between modalities. For each query, the modality achieving higher recall@5 against expected clusters is designated as the primary pipeline. The union output begins with the primary pipeline's full ranked list, followed by unique clusters from the secondary pipeline appended in their original rank order. This produces an untruncated result list (up to $2\times$ top-k), evaluated at recall@5, @10, @15, and @20. Ties at recall@5 are broken by mean reciprocal rank (MRR).

F.3. Analytical modules

F.3.1. CELL–CELL INTERACTION PREDICTION

ELISA predicts ligand–receptor (LR) interactions between cell types using a curated database of 280+ LR pairs spanning 25 signaling pathway categories. The database was compiled from established resources (CellChat(Jin et al., 2025), CellPhoneDB(Efremova et al., 2020), NicheNet(Browaeys et al., 2020)) and augmented with context-specific pairs for cystic fibrosis, neurodegeneration, neuroblastoma, and immune checkpoint biology. Each interaction is represented as a (ligand, receptor, pathway) tuple.

For each source–target cluster pair, the interaction score is computed as:

$$s_{ij} = \mathrm{pct}_{\mathrm{in}}(\text{ligand},\, c_i) \times \mathrm{pct}_{\mathrm{in}}(\text{receptor},\, c_j) \tag{8}$$

where $\mathrm{pct}_{\mathrm{in}}$ denotes the fraction of cells expressing the gene above detection threshold. Interactions are filtered by minimum expression thresholds (ligand $\geq 10\%$, receptor $\geq 5\%$ by default) and ranked by score. The module outputs per-interaction statistics, pathway-level summaries, and directional pair summaries.

F.3.2. PATHWAY ACTIVITY SCORING

Pathway activity across clusters is quantified using curated gene sets encompassing 60+ pathways organized into five categories: immune signaling (IFN-γ, Type I IFN, TNF/NF-κB, JAK-STAT, complement, TLR, chemokine), cell biology (mTOR, PI3K-Akt, Wnt, Notch, Hippo, Hedgehog, cell cycle, apoptosis), neuroscience (glutamatergic/GABAergic synapse,

neurodegeneration, FCD progenitor markers), metabolism (oxidative phosphorylation, glycolysis, lipid metabolism, fatty acid metabolism), and tissue-specific programs (surfactant metabolism, epithelial defense, fibrosis, angiogenesis).

For each pathway–cluster combination, the score is computed as the mean $\mathrm{pct}_{\mathrm{in}}$ (or alternative metric: $\log_2 \mathrm{FC}$, $\mathrm{pct}_{\mathrm{out}}$) across pathway genes detected in the cluster's DE profile, requiring a minimum of 3 genes for a non-zero score. Coverage (fraction of pathway genes detected) is reported alongside scores. Pathway query matching uses word-overlap fuzzy matching to accommodate variant pathway names.

F.3.3. Comparative analysis

When dataset metadata includes a condition column (e.g., "patient_group" with values "CF" and "Ctrl"), ELISA enables condition-stratified analysis. The module detects condition columns through keyword matching against a curated list (patient_group, condition, disease, treatment, genotype, *etc.*) and validates that the column contains 2–10 distinct values. For each cluster, the condition distribution is estimated from metadata field weights, and a condition bias label is assigned (>60% of cells from one condition). Per-gene statistics ($\log_2 \mathrm{FC}$, $\mathrm{pct}_{\mathrm{in}}$, $\mathrm{pct}_{\mathrm{out}}$) are reported within condition-biased clusters, and condition-enriched gene lists are compiled across all clusters.

F.3.4. Proportion analysis

The cell type proportion analysis computes the per-cluster cell counts and fractions relative to the total dataset size. When a condition column is available, the module additionally computes the condition-specific proportions and fold changes. For binary conditions (e.g., CF vs. Control), fold changes were calculated as the fraction of condition A cells in a cluster divided by the fraction of condition B cells, enabling the identification of cell types enriched or depleted in disease states.

F.3.5. Additional analytical functions

Supplementary analytical functions include: (i) marker specificity scoring, which ranks genes by a weighted score combining specificity $\left(\mathrm{pct}_{\mathrm{in}}/(\mathrm{pct}_{\mathrm{in}} + \mathrm{pct}_{\mathrm{out}})\right)$ and effect size ($|\log_2 \mathrm{FC}|$); (ii) co-expression analysis, computing Pearson correlations of $\mathrm{pct}_{\mathrm{in}}$ profiles across clusters; (iii) cell cycle scoring using established S-phase and G2M-phase gene signatures (43 and 54 genes, respectively); and (iv) gene set enrichment against 10 MSigDB Hallmark gene sets.

F.4. Visualization module

ELISA includes a comprehensive visualization module that generates publication-quality figures in two categories. Retrieval-level visualizations include: embedding landscape projections (UMAP(McInnes et al., 2018), t-SNE(Maaten & Hinton, 2008), or PCA fallback) of cluster-level semantic and expression embeddings, with optional highlighting of retrieved clusters; inter-cluster cosine similarity heatmaps; retrieval score waterfall plots; gene evidence bar charts ($\log_2 \mathrm{FC}$ or $\mathrm{pct}_{\mathrm{in}}$); gene-by-cluster heatmaps; radar charts for multi-metric cluster profiles; semantic vs. expression similarity scatter plots for hybrid retrieval diagnostics; and lambda sweep curves for fusion weight optimization.

When an AnnData (.h5ad) file is provided, cell-level visualizations are generated in a style consistent with Nature and Cell journals: cell-level UMAP plots with Cell Ontology labels placed using a centroid-offset algorithm with iterative repulsion to minimize label overlap; single-gene expression UMAPs with non-expressing cells shown in grey and expression on a purple gradient (capped at the 98th percentile); multi-gene expression grids; and dot plots showing percentage expression (dot size) and z-scored mean expression (dot color) across clusters. All plots used a 40-color colorblind-friendly palette and rasterized cell-level rendering for efficient file sizes.

F.5. LLM-mediated chat interface

The interactive chat interface wraps all modules behind a command-driven interface that routes user queries to the appropriate pipeline and generates LLM-interpreted summaries. The interface supports six retrieval and analysis modes (semantic, hybrid, discovery, compare, interactions, pathway, proportions) and 15 visualization commands. Each analysis result is automatically accumulated into a session-level report builder.

LLM interpretation is performed via the Groq API using the LLaMA-3.1-8B-Instant model(Grattafiori et al., 2024) at temperature 0.2. Prompts are constructed with mode-specific templates that enforce strict grounding in dataset evidence: the LLM receives only the retrieved cluster data, gene statistics, and pathway/interaction results as context, with explicit

instructions to avoid hallucination, external literature, and causal claims. Context payloads are trimmed to fit within the model's token limits (∼4,500 tokens for user content), with priority given to top-ranked clusters and highest-effect-size genes.

A discovery mode extends standard retrieval by prompting the LLM to produce four structured sections: (i) dataset evidence, (ii) established biology, (iii) consistency analysis identifying matches and mismatches with known biology, and (iv) candidate novel hypotheses stated with probabilistic language. This mode is designed to surface unexpected findings that may represent context-shifted gene functions or novel cell–cell interactions.

F.6. Benchmarking framework

F.6.1. Query design

The benchmark comprises 100 queries divided into two categories: 50 ontology queries (concept-level, testing semantic understanding) and 50 expression queries (gene-signature-based, testing transcriptomic matching). Queries were derived from the findings of Berg *et al.*(Berg et al., 2025), covering all major cell types identified in the study (macrophages, monocytes, $CD8^+$ T cells, $CD4^+$ T cells, B cells, basal cells, ciliated cells, NK cells, ionocytes, endothelial cells, dendritic cells, mast cells, secretory/goblet cells, fibroblasts, and neuroendocrine cells). Each query has a curated set of expected clusters and expected genes, enabling evaluation at both the cluster retrieval and gene delivery levels.

F.6.2. Baseline comparisons

ELISA's retrieval performance was evaluated against: the progression CellWhisperer, Semantic ELISA, scGPT ELISA, Additive Union.

F.6.3. Metrics

Retrieval performance was assessed using three metrics. **Cluster Recall@k** measures the fraction of expected clusters appearing in the top-k retrieved results, using fuzzy matching (substring containment or word-overlap Jaccard similarity ≥ 0.5) to accommodate Cell Ontology naming variations. **Mean Reciprocal Rank (MRR)** captures the rank position of the first relevant cluster. **Gene Recall** measures the fraction of expected genes recoverable from the DE profiles of the top-5 retrieved clusters, assessing whether retrieved clusters collectively provide the gene evidence needed for biological interpretation.

F.6.4. Analytical module evaluation

Analytical modules were evaluated against ground truth derived from the source publication. Interaction prediction was assessed by ligand–receptor pair recovery rate (whether the correct LR pair was detected regardless of cell type) and full match rate (correct LR pair between the correct source and target cell types, using fuzzy cell type matching). Pathway scoring was evaluated by alignment: the fraction of path activities reported on paper that ELISA correctly identified as active (score > 0) in at least one group. The proportion analysis was evaluated by the consistency rate whether cell types reported as increased or decreased in CF show fold changes in the expected direction. Comparative analysis was evaluated by gene recall, the fraction of differentially expressed genes reported on paper that can be recovered from the condition-stratified analysis of ELISA's.

F.7. Data representation and preprocessing

Each dataset is preprocessed into a single serialized PyTorch file (`.pt`) containing: cluster identifiers, precomputed BioBERT semantic embeddings (768-dimensional, L2-normalized), optional scGPT expression embeddings, per-cluster DE gene statistics ($\log_2 \mathrm{FC}$, $\mathrm{pct}_{\mathrm{in}}$, $\mathrm{pct}_{\mathrm{out}}$, adjusted p-value), per-cluster GO and Reactome enrichment terms, per-cluster metadata (cell counts, condition distributions, categorical field frequencies), cluster text descriptions, and the complete gene vocabulary. This representation enables ELISA to operate entirely at the cluster level without requiring access to the original count matrix, substantially reducing memory requirements and enabling deployment on standard hardware.

F.8. Software dependencies and reproducibility

ELISA depends on: PyTorch ($\geq$1.12) for tensor operations and data serialization, NumPy for numerical computation, sentence-transformers(Reimers & Gurevych, 2019) for BioBERT encoding, scikit-learn for t-SNE projections, UMAP-learn(McInnes et al., 2018) for UMAP projections, matplotlib for visualization, scanpy(Wolf et al., 2018) for AnnData-backed cell-level plots, SciPy for hierarchical clustering and sparse matrix operations, and the Groq Python SDK for LLM access. All analyses were performed on a standard workstation without GPU requirements for the retrieval and analytical modules; BioBERT (Lee et al., 2020) encoding benefits from but does not require GPU acceleration.

F.9. Future directions.

Several extensions can strengthen and broaden ELISA's capabilities. Integration with spatial transcriptomics data would enable spatially resolved interaction prediction, addressing the current limitation of expression-only interaction scoring. Incorporation of trajectory inference methods would allow ELISA to reason about dynamic processes such as differentiation and therapy response. Expansion of the retrieval engine to support cross-dataset queries comparing cell types across tissues or disease states would enable the kind of meta-analytical reasoning that was outside ELISA's scope in the ICB dataset evaluation. Finally, replacing the fixed LLM with a fine-tuned model trained on single-cell biological reasoning can improve the specificity and depth of automated interpretations.

F.10. ELISA parameters and hyperparameters

Tables 6–9 report all parameters and hyperparameters used in the ELISA framework. Default values were used throughout all experiments; no dataset-specific tuning was performed.

Table 6. **Data preprocessing and embedding generation parameters.**

Parameter	Value	Description
Preprocessing (Scanpy)		
`target_sum`	10,000	Library-size normalization target
`n_top_genes`	3,000	HVGs selected (Seurat v3)
`max_value`	10	Z-score clipping threshold
`n_comps`	50	PCA components
Leiden resolution	1.0	Used only if no annotations exist
Differential expression		
Method	Wilcoxon	Via `scanpy.tl.rank_genes_groups`
`DE_PVAL`	0.10	Adjusted p-value cutoff
`TOP_K_MARKERS_STATS`	10,000	Max genes stored per cluster
`TOP_K_MARKERS_TEXT`	400	Genes in cluster text summaries
Enrichment (gseapy)		
Gene sets	GO_Biological_Process_2023, Reactome_2022	
`TOP_K_GO`	15	GO terms retained per cluster
`TOP_K_REACTOME`	15	Reactome terms per cluster
Enrichment cutoff	0.05	Adjusted p-value threshold
Input genes	200	Top DE genes per enrichment call
Semantic embedding (BioBERT)		
Model	`pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb`	
Embedding dim	768	Output dimensionality
α (`IDENTITY_ALPHA`)	0.6	Identity vs. context weight
Normalization	L2	Final combined embeddings
Batch size	16	Sentences per encoding batch
scGPT expression embedding		
Model	scGPT whole-human	Pre-trained foundation model
Embedding dim	512	CLS token dimensionality
`N_BINS`	51	Expression binning resolution
`MAX_TOKENS`	3,000	Max gene tokens per cell
Batch size	64	Cells per inference batch
Aggregation	Mean pooling	Cell $\rightarrow$ cluster centroids
Normalization	L2	Cluster-level centroids

Table 7. **Hybrid retrieval engine parameters.**

Parameter	Value	Description
Query classification		
Gene threshold	$\geq$60%	Token fraction to route as gene query
Mixed threshold	$\geq$20% each	Gene + NL tokens for mixed routing
Gene pattern	A–Z, 2–15 chars	Regex for gene symbol detection
Gene marker scoring		
Score function	$(0.5 + \lvert\log_2 \text{FC}\rvert) \times (0.3 + \max(\text{pct}_{\text{in}} - \text{pct}_{\text{out}}, 0))$	
High-expr bonus	$\times 1.3$	When $\text{pct}_{\text{in}} > 0.5$
Coverage factor	$0.5 + 0.5 \times n_{\text{found}}/n_{\text{query}}$	
Semantic matching		
Similarity	Cosine	Query vs. cluster embeddings
Name boost (α)	0.15	Bonus for ontology name overlap
Min substring	4 chars	For name boost activation
Synonym boost (β)	0.10	Bonus for synonym match
Reciprocal rank fusion		
RRF constant (k)	60	Smoothing constant
Weights	1.0 : 1.0	Gene : semantic
Additive union (benchmarking)		
Primary selection	Recall@5	Higher-recall modality is primary
Tiebreaker	MRR	When Recall@5 is tied
Default settings		
`top_k`	5	Clusters returned per query
`pre_k`	40	Candidates before reranking
γ	2.5	Reranking sharpness
λ_{sem} (scGPT)	0.0	Pure gene scoring mode
λ_{sem} (discovery)	0.5	Balanced mode

Table 8. **Analytical module parameters.**

Parameter	Value	Description
Ligand–receptor interactions		
Database size	280+ pairs	From CellChat, CellPhoneDB, NicheNet
Pathway categories	25	Signaling annotations
`min_ligand_pct`	0.10	Min ligand expr. in source
`min_receptor_pct`	0.05	Min receptor expr. in target
Score	$\text{pct}_{\text{in}}(L) \times \text{pct}_{\text{in}}(R)$	Expression fraction product
Self-interactions	Excluded	Source $\neq$ target
Pathway activity scoring		
Number of pathways	60+	Across 5 categories
Metric	Mean pct_{in}	Avg. expression of pathway genes
`min_genes`	3	Min for non-zero score
Categories	Immune, Cell biology, Neuroscience, Metabolism, Tissue-specific	
Comparative analysis		
Condition bias	$>60\%$	Fraction to assign bias label
`min_pct`	0.05	Min expr. for gene inclusion
`top_n`	20	Genes per cluster
Enriched genes	30	Per-condition summary limit
Proportion analysis		
Fold change	$\text{frac}_A/\text{frac}_B$	Condition ratio
Min denominator	0.001	Below: reported as ∞
Cell cycle scoring		
S-phase genes	43	Seurat S-phase markers
G2M-phase genes	54	Seurat G2M markers
Cycling threshold	$S>0.3$ and G2M>0.3	Both above threshold
Gene set enrichment		
Default gene sets	10 MSigDB Hallmark	Curated pathways
`min_genes`	3	Min for non-zero score

Table 9. **LLM interpretation parameters.**

Parameter	Value	Description
LLM configuration		
Default provider	Groq	Free tier, 500K tokens/day
Default model	LLaMA-3.1-8B	Via Groq Cloud API
Supported	4 providers	Groq, Gemini, OpenAI, Claude
Temperature	0.2	Low for reproducibility
Prompt limit	18,000 chars	$\approx$4,500 tokens
Context limit	12,000 chars	$\approx$3,000 tokens
Safety and rate limiting		
Spending cap	€1.00	Hard cap, configurable
Max retries	5	On rate-limit errors
Initial wait	10 s	Backoff start
Backoff	Exponential	Max 120 s
Context trimming		
Clusters	Top 10	In compare mode
Gene evidence	Top 5	Per cluster
Pathway scores	Top 10	Entries to LLM
Interactions	Top 20	Entries to LLM
Discovery sections	4	Evidence, Biology, Consistency, Hypotheses

F.11. D1: Cystic Fibrosis Airways ((Berg et al., 2025) *et al.*)

F.11.1. Ontology Queries

Q1. Macrophage and monocyte infiltration in cystic fibrosis airways

Q2. Recruited monocytes and pro-inflammatory macrophages in CF lung tissue

Q3. Macrophage scavenging receptor expression and phagocytosis in CF

Q4. Non-classical monocyte patrol function in CF bronchial wall

Q5. CD8 T cell activation and cytotoxicity in CF lung inflammation

Q6. CD8 T cell inflammatory cytokine production and IFNG signaling in CF

Q7. HLA-E CD94 NKG2A immune checkpoint inhibiting CD8 T cell activity

Q8. Dysfunctional CD8 T cell response to chronic *Pseudomonas* infection in CF

Q9. CALR LRP1 interaction between T cells and macrophages promoting inflammation

Q10. CD4 helper T cell immune activation in cystic fibrosis

Q11. CD4 T cell VEGF receptor signaling and hypoxia response in CF

Q12. Aberrant Th2 and Th17 T cell responses in *Pseudomonas*-infected CF lungs

Q13. Chronic adaptive immune activation of T lymphocytes in CF despite modulator therapy

Q14. B cell activation and immunoglobulin response in CF airways

Q15. B cell receptor downregulation and reduced plasma cell markers in CF

Q16. Interferon gamma signaling and HLA-DP expression in B cells of CF patients

Q17. PDGFRB signaling pathway activated in B cells from CF lungs

Q18. Basal cell dysfunction and reduced stemness in cystic fibrosis epithelium

Q19. Impaired basal cell differentiation and pathogenic basal cell variants in CF

Q20. Basal cell DNA damage repair and chromatin remodeling in CF airways

Q21. Reduced keratinization gene expression CSTA HSPB1 in CF basal cells

Q22. Basal cell altered cell–cell communication and increased interactions in CF

Q23. Ciliated cell ciliogenesis and increased abundance in CF bronchial epithelium

Q24. Ciliated cell HLA class II expression and immune-linked transcriptional changes in CF

Q25. Skewed basal cell differentiation towards ciliated cells in CF epithelium

Q26. Natural killer cell cytotoxicity and NKG2A immune checkpoint in CF

Q27. NKG2A blockade to restore NK and CD8 T cell function in CF lung

Q28. Innate lymphoid cell dysfunction and impaired antimicrobial defense in CF

Q29. Pulmonary ionocyte CFTR expression in cystic fibrosis

Q30. Ionocyte unique cell–cell interactions with adaptive lymphocytes in CF

Q31. Endothelial cell remodeling and VEGF signaling in CF lung

Q32. Reduced endothelial cell proportions and altered differentiation in CF airways

Q33. Hypoxia-induced VEGF upregulation and vascular remodeling in CF lungs

Q34. Dendritic cell antigen presentation in CF airways

Q35. IFNG IFNGR2 interaction between CD8 T cells and dendritic cells in CF

Q36. Mast cell degranulation and allergic inflammation in CF

Q37. Secretory cell mucus overproduction and inflammatory signaling in CF epithelium

Q38. Goblet cell hyperplasia and mucin gene expression in cystic fibrosis

Q39. Submucosal gland epithelial cell changes in cystic fibrosis

Q40. Reduced submucosal gland cell proportions and gland development dysfunction in CF

Q41. Type I interferon response and inflammatory signaling in CF epithelial cells

Q42. Interferon responsive gene upregulation across epithelial subsets in CF

Q43. VEGF receptor signaling and hypoxia response across cell types in CF

Q44. TXNIP-mediated NLRP3 inflammasome activation in CF lymphocytes and epithelial cells

Q45. GNAI2 immunomodulatory signaling in CD8 T cells and B cells in CF

Q46. GNAI2 adenylate cyclase regulation and CFTR function in lymphocytes

Q47. Stromal cell and fibroblast remodeling in CF airway tissue

Q48. Pericyte and stromal cell contribution to airway fibrosis in CF

Q49. IFNG–IFNGR1 interaction between CD8 T cells and basal cells, macrophages, and endothelial cells in CF

Q50. Altered structural–immune cell crosstalk in CF involving lymphocytes, ionocytes, and macrophages

F.11.2. Expression Queries

Q51. `MARCO FABP4 APOC1 C1QB C1QC MSR1`

Q52. `CD68 CD14 CSF1R CSF2RA LGALS2`

Q53. `GOS2 FABP4 PPARG APOC1 C1QB`

Q54. `FCGR3A CX3CR1 CD14 CDKN1C LILRB2`

Q55. `CD8A CD8B GZMB PRF1 IFNG NKG7`

Q56. `IFNG GNAI2 CD69 CD81 CD3G FOS JUND`

Q57. `GZMB PRF1 NKG7 GNLY KLRD1 CD8A`

Q58. `TXNIP MAP2K2 IFNG CD81 CD3G CD69`

Q59. `KLF2 IL7R CD48 TXNIP ETS1`

Q60. `CD3D CD4 IL7R CD3E CD3G`

Q61. `TRAJ52 TRBV22-1 TRDJ2 CD3E CD3G`

Q62. `CD3G CD3E CD69 IL7R CD81 FOS`

Q63. `IGLJ3 IGKJ1 IGHJ5 JCHAIN MZB1 XBP1`

Q64. CD79A IGHG3 IGLC2 SYK CD81 JCHAIN

Q65. SYK CSK CD9 CD81 JUND LTB HLA-DPA1

Q66. IGHG3 IGLC2 IGHD IGHA1 IGLC1 IGLC3

Q67. KRT5 KRT14 KRT15 TP63 IL33 CSTA

Q68. CSTA HSPB1 KRT5 KRT14 TP63

Q69. KRT5 IL33 TP63 KRT15 LAMB3 COL17A1

Q70. FOXJ1 DNAH5 CAPS PIFO RSPH1 DNAI1

Q71. DNAH5 SYNE1 SYNE2 CAPS PIFO

Q72. GNLY KLRD1 KLRK1 NKG7 PRF1 GZMB

Q73. GNLY NKG7 KLRD1 KLRK1 KLRC1

Q74. ATP6V1G3 FOXI1 BSND CLCNKB ASCL3

Q75. FOXI1 CFTR ATP6V1G3 BSND RARRES2

Q76. PLVAP ACKR1 ERG VWF PECAM1 CDH5

Q77. VIM PLVAP ACKR1 MGP PTGDS CXCL14

Q78. CPA3 TPSAB1 TPSB2 MS4A2 HDC GATA2

Q79. TPSAB1 TPSB2 KIT CPA3 MS4A2

Q80. HLA-DPA1 HLA-DRB1 CD74 GPR183 LGALS2

Q81. HLA-DPA1 HLA-DPB1 HLA-DRB1 CD80 CD86 CD74

Q82. SCGB1A1 SCGB3A1 MUC5AC MUC5B LYPD2 PRR4

Q83. SCGB1A1 MUC5AC SCGB3A1 LYPD2

Q84. MUC5AC MUC5B LYZ SCGB1A1 SCGB3A1

Q85. COL1A2 LUM DCN SFRP2 COL3A1 PDGFRA

Q86. PDGFRA COL1A2 COL3A1 VCAN DCN LUM

Q87. PDGFRB VIM COL1A2 MGP CXCL14

Q88. SST CHGA ASCL1 GRP CALCA SYP

Q89. GRP ASCL1 SYT1 CHGA SYP CALCA

Q90. HLA-E KLRC1 KLRD1 KLRC2 KLRC3 KLRK1

Q91. HLA-E KLRC1 KLRD1 CD8A CD8B

Q92. CALR LRP1 GNAI2 FOS JUND MAP2K2

Q93. GNAI2 CXCR3 F2R S1PR4 CD69

Q94. IFIT1 MX1 OAS2 ISG15 IFITM3 IFIT3

Q95. IFIT1 MX1 OAS2 IFIT3 IFI6

Q96. KDM1A KMT5A RAD50 ERCC6 ERCC8

Q97. TXNIP MAP2K2 ETS1 VEGFA KLF2

Q98. IFNG IFNGR1 IFNGR2 CALR LRP1

Q99. CCL5 CCR5 CXCL10 CXCR3 F2R

Q100. CFTR FOXI1 SCGB1A1 KRT5 FOXJ1 MUC5AC

F.12. D5: Healthy Breast Tissue Atlas ((Bhat-Nakshatri et al., 2024) *et al.*)

F.12.1. Ontology Queries

Q1. Luminal hormone sensing cells with estrogen receptor expression in the healthy breast

Q2. FOXA1 pioneer transcription factor activity in luminal hormone responsive breast epithelial cells

Q3. ERα–FOXA1–GATA3 transcription factor network in hormone responsive breast cells

Q4. Mature luminal cells with hormone receptor positive identity in breast tissue

Q5. Hormone sensing alpha versus beta cell states in breast epithelium

Q6. LHS cell-enriched fate factor DACH1 and PI3K pathway regulator INPP4B in breast

Q7. Lobular epithelial cells expressing APOD and immunoglobulin genes in breast

Q8. Luminal adaptive secretory precursor cells and progenitor identity in breast

Q9. ELF5 and EHF transcription factor expression in luminal progenitor breast cells

Q10. Alveolar progenitor cell state enriched in Indigenous American breast tissue

Q11. BRCA1 associated breast cancer originating from luminal progenitor cells

Q12. KIT receptor expression and chromatin accessibility in luminal progenitor cells

Q13. MFGE8 and SHANK2 expression in luminal progenitor cells of the breast

Q14. LASP basal–luminal intermediate progenitor cell identity in the breast

Q15. Basal-myoepithelial cells with TP63 and KRT14 expression in breast

Q16. Basal cell chromatin accessibility and TP63 binding site enrichment

Q17. Basal alpha and basal beta cell states in breast myoepithelium

Q18. SOX10 motif enrichment in basal-myoepithelial cells of the breast

Q19. KRT14 KRT17 expression in ductal epithelial and basal cells of breast tissue

Q20. Fibroblast heterogeneity and cell states in healthy breast stroma

Q21. Genetic ancestry-dependent variability in breast fibroblast cell states

Q22. Fibro-prematrix state enrichment in African ancestry breast tissue fibroblasts

Q23. PROCR ZEB1 PDGFRα multipotent stromal cells enriched in African ancestry breast

Q24. Myofibroblast and inflammatory fibroblast subtypes in breast cancer stroma

Q25. SFRP4 and Wnt pathway modulation in breast fibroblasts

Q26. Endothelial cell subtypes and vascular markers in breast tissue

Q27. Lymphatic endothelial cells expressing LYVE1 in breast stroma

Q28. ACKR1 stalk-like endothelial cell subtype in breast vasculature

Q29. Vascular endothelial cell heterogeneity in mammary gland microvasculature

Q30. Breast tissue angiogenesis and endothelial cell MECOM expression

Q31. T lymphocyte markers and immune cell identity in breast tissue

Q32. CD4 T cell IL7R expression and chromatin accessibility in breast

Q33. CD8 T cell GZMK cytotoxic activity and IFNG signaling in breast tissue

Q34. Tissue-resident memory T lymphocyte populations in healthy breast

Q35. Adaptive immune surveillance by T cells in mammary gland stroma

Q36. Macrophage identity and FCGR3A expression in breast tissue stroma

Q37. Macrophage subtypes and tissue-resident immune cells in healthy breast

Q38. Breast tissue-resident macrophage phagocytic function and complement expression

Q39. Myeloid lineage immune cells and monocyte-derived macrophages in mammary gland

Q40. Adipocyte subtypes and lipid metabolism in breast tissue

Q41. Adipocyte PLIN1 and FABP4 expression in healthy breast stroma

Q42. PLIN1 lipid droplet biology and adipocyte identity in mammary fat pad

Q43. Mammary gland adipose tissue and fatty acid binding protein expression

Q44. Epithelial cell hierarchy from basal to luminal hormone sensing in breast

Q45. CXCL12 chemokine expression in endothelial cells and fibroblasts of breast

Q46. VEGFA angiogenic signaling from luminal cells to endothelium in breast

Q47. IGF1 paracrine signaling from fibroblasts to luminal cells in breast stroma

Q48. Breast tissue microenvironment with stromal and immune cell interactions

Q49. Ancestry differences in breast tissue cellular composition and cancer risk

Q50. Gene expression differences between ductal and lobular epithelial cells of the breast

F.12.2. Expression Queries

Q51. `FOXA1 ESR1 GATA3 ERBB4 ANKRD30A AFF3 TTC6`

Q52. `MYBPC1 THSD4 CTNND2 DACH1 INPP4B NEK10`

Q53. `ESR1 FOXA1 GATA3 ELOVL5 ANKRD30A`

Q54. `AFF3 TTC6 ERBB4 MYBPC1 THSD4`

Q55. `DACH1 NEK10 CTNND2 INPP4B ELOVL5`

Q56. `APOD IGHA1 IGKC ESR1 FOXA1 GATA3`

Q57. `DUSP1 DPM3 RPL36 IGHA1 IGKC APOD`

Q58. `ELF5 EHF KIT CCL28 KRT15 BARX2 NCALD`

Q59. `MFGE8 SHANK2 SORBS2 AGAP1 ELF5`

Q60. KRT15 CCL28 KIT INPP4B ELF5

Q61. RBMS3 EHF BARX2 NCALD ELF5

Q62. ESR1 ELF5 EHF KIT CCL28

Q63. ELF5 KIT CCL28 EHF KRT15 BARX2

Q64. NCALD BARX2 SHANK2 SORBS2 MFGE8 ELF5

Q65. TP63 KRT14 KLHL29 FHOD3 SEMA5A

Q66. KLHL13 KLHL29 TP63 KRT14 PTPRT

Q67. TP63 KRT14 KRT17 FHOD3 ABLIM3

Q68. ST6GALNAC3 PTPRM SEMA5A KLHL29

Q69. KRT14 KRT17 TP63 KLHL29 KLHL13 FHOD3

Q70. LAMA2 SLIT2 RUNX1T1 COL1A1 COL3A1

Q71. COL3A1 POSTN COL1A1 IGF1 ADAM12

Q72. CFD MGST1 MFAP5 COL3A1 POSTN

Q73. PROCR ZEB1 PDGFRA COL1A1 LAMA2

Q74. SFRP4 COL1A1 POSTN LAMA2 SLIT2

Q75. COL1A1 PDPN CD34 CXCL12 LAMA2

Q76. MECOM LDB2 MMRN1 CXCL12 ACKR1

Q77. LYVE1 MECOM LDB2 MMRN1

Q78. ACKR1 CXCL12 MECOM LDB2

Q79. MECOM LDB2 MMRN1 LYVE1 ACKR1

Q80. CXCL12 MECOM LDB2 ACKR1 MMRN1

Q81. PTPRC SKAP1 ARHGAP15 THEMIS IL7R

Q82. IL7R GZMK PTPRC SKAP1

Q83. IFNG GZMK IL7R THEMIS PTPRC

Q84. THEMIS ARHGAP15 SKAP1 PTPRC IL7R

Q85. PTPRC SKAP1 GZMK IFNG THEMIS ARHGAP15

Q86. FCGR3A ALCAM LYVE1 CD163

Q87. ALCAM FCGR3A LYVE1 CD14

Q88. FCGR3A ALCAM CD163 MERTK

Q89. ALCAM LYVE1 FCGR3A CD163 MARCO

Q90. PLIN1 FABP4 KIT ADIPOQ LEP

Q91. FABP4 PLIN1 ADIPOQ LEP LPL

Q92. PLIN1 FABP4 LPL PPARG ADIPOQ

Q93. FABP4 PLIN1 KIT ADIPOQ

Q94. FOXA1 ELF5 TP63 KRT14 GATA3 ESR1

Q95. GATA3 EHF ELF5 FOXA1 KRT15 KRT14 TP63

Q96. MECOM PTPRC FCGR3A PLIN1 LAMA2 TP63 FOXA1

Q97. CXCL12 LAMA2 MECOM LDB2 COL1A1

Q98. ESR1 FOXA1 ELF5 EHF KIT TP63 KRT14

Q99. PTPRC FCGR3A FABP4 PLIN1 MECOM

Q100. VEGFA LDB2 IGF1 LAMA2 FOXA1 ELF5

F.13. D3: Fetal Lung AT2 Organoids ((Lim et al., 2025) *et al.*)

F.13.1. Ontology Queries

Q1. Alveolar type 2 cell identity and surfactant protein production in fetal lung organoids

Q2. Mature AT2 cell markers and lamellar body formation in fdAT2 organoids

Q3. Surfactant protein C maturation and intracellular trafficking in alveolar epithelium

Q4. SFTPC processing through endosomal compartments and multivesicular bodies

Q5. Surfactant secretion and lamellar body exocytosis in human AT2 cells

Q6. ITCH E3 ubiquitin ligase role in SFTPC trafficking and ubiquitination

Q7. K63 ubiquitination of surfactant protein C for ESCRT recognition and MVB entry

Q8. HECT domain E3 ligase ITCH depletion phenocopying SFTPC-I73T pathogenic variant

Q9. Ubiquitome forward genetic screen for SFTPC trafficking effectors

Q10. SFTPC relocalisation to plasma membrane and recycling endosomes upon ITCH loss

Q11. AT2 stem cell self-renewal and proliferation in fetal lung organoids

Q12. FGF7-driven AT2 cell proliferation and surfactant processing balance

Q13. Expandable fetal-derived AT2 organoids maintaining identity over passaging

Q14. Alveolar type 1 cell differentiation from AT2 organoids via YAP activation

Q15. AT2 to AT1 lineage transition through Wnt withdrawal and LATS inhibition

Q16. AT1 cell fate markers AQP5 CAV1 AGER in differentiated fdAT2 organoids

Q17. CXCL chemokine expressing AT2 subpopulation in fetal lung organoids

Q18. Immune response gene expression in alveolar type 2 cells

Q19. Chemokine-mediated innate immune signaling in AT2 organoid subsets

Q20. Aberrant basal cell differentiation from AT2 cells in organoid culture

Q21. Hypoxia-induced airway differentiation of alveolar type 2 cells

Q22. Pulmonary neuroendocrine cell differentiation in AT2 organoids

Q23. Neuroendocrine progenitor cells co-expressing SFTPC and NE markers

Q24. Ciliated cell-like differentiation in fetal AT2 organoid culture

Q25. Intermediate transitional cell state between AT2 and differentiated lineages

Q26. Surfactant metabolism and lipid transport in fetal alveolar epithelium

Q27. Vesicle-mediated transport and lysosome localization in AT2 surfactant processing

Q28. Lipid storage membrane transport and vesicle cytoskeleton trafficking in AT2 cells

Q29. Wnt signaling pathway maintaining AT2 identity and inhibiting AT1 differentiation

Q30. SFTPC-I73T pathogenic variant causing interstitial lung disease and AT2 dysfunction

Q31. Toxic gain-of-function effect of misfolded surfactant protein C variants

Q32. Transcriptional maturity of fdAT2 organoids compared to adult AT2 and PSC-iAT2

Q33. Missing immune response MHC class II genes in fetal versus adult AT2 cells

Q34. CRISPRi-mediated depletion of ITCH and UBE2N in fdAT2 organoids

Q35. Reversible SFTPC mislocalization after CRISPRi recovery in AT2 organoids

Q36. ESCRT complex components HRS VPS28 required for SFTPC MVB entry

Q37. Endosomal recycling of SFTPC to plasma membrane upon ubiquitination failure

Q38. SUMOylation pathway components UBE2I UBA2 PIAS1 and SFTPC expression regulation

Q39. Fetal lung tip progenitor differentiation into mature AT2 cells

Q40. EpCAM positive tip epithelial cell isolation and AT2 organoid derivation

Q41. SFTPC C-terminal cleavage and proprotein processing in endosomal compartments

Q42. proSFTPC plasma membrane transit before endocytosis and maturation

Q43. Interstitial lung disease caused by SFTPC variants and AT2 cell dysfunction

Q44. Heritable pulmonary fibrosis from SFTPC mistrafficking and toxic accumulation

Q45. AT2 medium components dexamethasone cAMP IBMX DAPT for alveolar differentiation

Q46. fdAT2 organoid engraftment in mouse precision-cut lung slices and AT1 differentiation

Q47. NEDD4-2 HECT domain ligase role in SFTPC ubiquitination and maturation

Q48. Cell type heterogeneity and proportions across fdAT2 organoid lines

Q49. fdAT2 organoid stability over long-term passaging and cryopreservation

Q50. Genetic manipulation of fetal AT2 organoids using lentiviral CRISPRi system

F.13.2. EXPRESSION QUERIES

Q51. SFTPC SFTPB SFTPA1 SFTPA2 NAPSA LAMP3

Q52. SFTPC SFTPB ABCA3 LAMP3 HOPX NKX2-1

Q53. NKX2-1 SLC34A2 LPCAT1 HOPX CEACAM6

Q54. SFTPC SFTPD SFTA3 CD36 CAV1 SLC34A2

Q55. SFTPA1 SFTPA2 SFTPB SFTPC SFTPD

Q56. ITCH UBE2N HRS VPS28 RABGEF1 EEA1

Q57. ITCH NEDD4 NEDD4L UBE2N UBE2I

Q58. EEA1 MICALL1 LAMP3 HRS VPS28

Q59. UBE2I UBA2 PIAS1 ITCH RABGEF1

Q60. ABCA3 LAMP3 NAPSA CKAP4 ZDHHC2 CTSH

Q61. ABCA3 SFTPB SFTPC LAMP3 P2RY2 LMCD1

Q62. MKI67 PCNA TOP2A SFTPC NKX2-1

Q63. MKI67 PCNA CDK1 CCNB1 SFTPC

Q64. CXCL1 CXCL2 CXCL3 CCL2 SFTPC

Q65. CXCL1 CXCL3 CCL2 CCL4 CCL4L1

Q66. CXCL1 CXCL2 HLA-DPA1 HLA-DPB1 CCL2

Q67. HLA-DQB1 HLA-DMA HLA-DMB HLA-DRA HLA-DOA

Q68. HLA-DPA1 HLA-DPB1 HLA-DRA CD86 TNF

Q69. AQP5 CAV1 AGER HOPX

Q70. CAV1 AGER AQP5 PDPN

Q71. TP63 KRT5 KRT14 KRT15 SOX2

Q72. KRT5 KRT14 TP63 LAMB3 COL17A1

Q73. ASCL1 NEUROD1 GRP CHGA SYP CALCA

Q74. GRP ASCL1 SYT1 CHGA SYP

Q75. ASCL1 GRP SFTPC NKX2-1

Q76. FOXJ1 DNAH5 CAPS PIFO RSPH1

Q77. FOXJ1 DNAH5 DNAI1 RSPH1 CAPS

Q78. SOX2 SOX9 NKX2-1 SFTPC TP63

Q79. SOX2 NKX2-1 HOPX CAV1

Q80. CTNNB1 TCF7L2 AXIN2 WNT3A LGR5

Q81. SFTPC NKX2-1 HOPX SFTPB ABCA3 MKI67

Q82. NAPSA ABCA3 SFTA3 SFTPD LAMP3 HOPX

Q83. SFTPC ITCH EEA1 LAMP3 MICALL1 ABCA3

Q84. SFTPC NAPSA CTSH LAMP3 ITCH UBE2N

Q85. SFTPC CXCL1 CXCL2 NKX2-1 LAMP3

Q86. CDH1 TJP1 EPCAM SFTPC NKX2-1

Q87. ITCH HRS VPS28 UBE2N RABGEF1 PIAS1 UBE2I UBA2

Q88. ITCH NEDD4 NEDD4L HRS UBAP1 USP8

Q89. MKI67 TOP2A PCNA CDK1 CCNB1 CCNA2

Q90. SFTPC TP63 ASCL1 FOXJ1 NKX2-1

Q91. SFTPC SFTPB ASCL1 GRP TP63 KRT5

Q92. SFTPC CAV1 AGER AQP5 HOPX NKX2-1

Q93. LAMP3 ABCA3 SFTPB SFTPC NAPSA CD36

Q94. CKAP4 ZDHHC2 SLC34A2 CTSH SFTPC

Q95. CXCL1 CXCL2 CXCL3 CCL2 CCL4 TNF

Q96. SOX9 NKX2-1 SFTPC SFTPB LAMP3

Q97. SFTPC NKX2-1 ASCL1 NEUROD1 GRP MKI67

Q98. SFTA3 SFTPD NAPSA NKX2-1 CKAP4 ZDHHC2 SLC34A2 CTSH SFTPA1 SFTPA2 SFTPC SFTPB

Q99. ITCH SFTPC LAMP3 ABCA3 UBE2N NAPSA

Q100. SFTPC CXCL1 MKI67 TP63 ASCL1 FOXJ1 SOX2 CAV1

F.14. D2: High-Risk Neuroblastoma ((Yu et al., 2025) *et al.*)

F.14.1. Ontology Queries

Q1. Neuroblast neoplastic cell of sympathetic nervous system expressing PHOX2B and ISL1

Q2. Neuroblastoma tumor cell with MYCN amplification and proliferative phenotype

Q3. Adrenergic neuroblast expressing catecholamine biosynthesis enzymes tyrosine hydroxylase

Q4. Neuroblastoma cell with calcium and synaptic signaling pathway enrichment

Q5. Dopaminergic neuroblast expressing dopamine transporter and metabolic genes

Q6. Proliferating neuroblastoma cell with cell cycle and DNA replication markers

Q7. Mesenchymal neuroblastoma cell state expressing extracellular matrix genes and YAP1

Q8. Intermediate OXPHOS neuroblast with ribosomal gene expression and oxidative phosphorylation

Q9. EZH2 expressing neuroblastoma cell PRC2 polycomb repressive complex chromatin regulation

Q10. Neuroblastoma cell ERBB4 receptor expressing epidermal growth factor signaling

Q11. Neuroblast with adrenergic transcription factor PHOX2A PHOX2B GATA3 expression

Q12. Neural crest derived neoplastic cell in pediatric tumor expressing chromogranin

Q13. Neuroblastoma cell immune evasion NECTIN2 and checkpoint ligand expression

Q14. Mesenchymal transition state in neuroblastoma with AP-1 transcription factors

Q15. Tumor associated macrophage in neuroblastoma microenvironment CD68 CD163 expressing

Q16. Pro-inflammatory macrophage IL18 expressing anti-tumor immune response

Q17. Pro-angiogenic macrophage VCAN expressing promoting tumor vascularization

Q18. Immunosuppressive macrophage C1QC SPP1 complement expressing in tumor

Q19. Tissue resident macrophage F13A1 expressing phagocytic function in neuroblastoma

Q20. Lipid associated macrophage HS3ST2 with metabolic phenotype in tumor

Q21. Macrophage secreting HB-EGF ligand for ERBB4 receptor activation on neuroblasts

Q22. CCL4 expressing pro-angiogenic macrophage chemokine signaling in tumor

Q23. Proliferating macrophage MKI67 TOP2A expanding after chemotherapy

Q24. THY1 positive macrophage undefined myeloid phenotype in neuroblastoma

Q25. T cell lymphocyte infiltrating neuroblastoma tumor expressing CD247 CD96

Q26. Cytotoxic T cell with granzyme perforin mediated tumor cell killing

Q27. Tumor infiltrating T lymphocyte immune response to neuroblastoma

Q28. B cell lymphocyte PAX5 MS4A1 in neuroblastoma tumor immune microenvironment

Q29. B lymphocyte humoral immunity and antigen presentation in pediatric tumor

Q30. Dendritic cell IRF8 FLT3 antigen presentation priming T cell responses in tumor

Q31. Professional antigen presenting dendritic cell MHC class II expression

Q32. Fibroblast stromal cell PDGFRB DCN extracellular matrix production in neuroblastoma

Q33. Cancer associated fibroblast FAP ACTA2 expressing in tumor stroma

Q34. Neural crest derived endoneurial fibroblast in neuroblastoma tissue

Q35. Schwann cell PLP1 CDH19 myelinating glial cell in neuroblastoma microenvironment

Q36. Schwann cell precursor neural crest lineage expanding after therapy

Q37. Endothelial cell PECAM1 PTPRB vascular marker in neuroblastoma tumor vasculature

Q38. Tumor endothelium blood vessel lining cell expressing vascular endothelial markers

Q39. Adrenal cortex cell steroidogenesis CYP11A1 CYP11B1 adjacent normal tissue

Q40. Cortical cell of adrenal gland steroid hormone biosynthesis normal adjacent tissue

Q41. Hepatocyte ALB expressing liver cell from adjacent normal tissue in neuroblastoma biopsy

Q42. Kidney cell renal tissue PKHD1 from adjacent normal tissue in neuroblastoma specimen

Q43. Chemotherapy induced tumor microenvironment rewiring macrophage expansion after therapy

Q44. HB-EGF ERBB4 paracrine signaling axis between macrophage and neuroblast promoting ERK

Q45. Tumor immune evasion and antigen presentation in neuroblastoma

Q46. VEGFA angiogenesis signaling in neuroblastoma tumor microenvironment

Q47. Immune cell infiltration in high-risk neuroblastoma T cell B cell macrophage

Q48. THBS1 CD47 don't eat me signal between macrophage and neuroblastoma cell

Q49. Neuroblastoma cell expressing ALK receptor tyrosine kinase oncogenic driver

Q50. Tumor microenvironment cell diversity neuroblasts fibroblasts Schwann endothelial macrophages

F.14.2. Expression Queries

Q51. `PHOX2B ISL1 HAND2 TH DBH DDC CHGA`

Q52. `MYCN MKI67 TOP2A EZH2 SMC4 BIRC5`

Q53. `PHOX2A PHOX2B GATA3 ASCL1 ISL1 HAND2`

Q54. `CACNA1B SYN2 KCNMA1 KCNQ3 GPC5 CREB5`

Q55. `SLC18A2 TH DDC AGTR2 ATP2A2 PHOX2B`

Q56. `MKI67 TOP2A EZH2 SMC4 BIRC5 BUB1B ASPM KIF11`

Q57. `YAP1 FN1 VIM COL1A1 SERPINE1 SPARC THBS2`

Q58. `ERBB4 EGFR HBEGF TGFA EREG AREG`

Q59. `NECTIN2 CD274 B2M HLA-A HLA-B PHOX2B`

Q60. `JUN FOS JUNB JUND FOSL2 BACH1 BACH2`

Q61. `CHGA CHGB PHOX2B ISL1 NTRK1 RET`

Q62. `ETS1 ETV6 ELF1 KLF6 KLF7 RUNX1 ZNF148`

Q63. `ALK MYCN NTRK2 PHOX2B TH`

Q64. `CD68 CD163 CD86 CSF1R MRC1 SPP1`

Q65. `IL18 CD68 CD163 CD86 HLA-DRA CSF1R`

Q66. `VCAN VEGFA CD68 CD163 SPP1 EGFR`

Q67. `C1QC SPP1 CD68 CD163 APOE TREM2`

Q68. `F13A1 CD68 CD163 MRC1 LYVE1 CSF1R`

Q69. `HS3ST2 CYP27A1 CD68 CD163 APOE LPL`

Q70. `HBEGF TGFA EREG AREG CD68 CD163`

Q71. `CCL4 CD68 CD163 VEGFA CSF1R CCL3`

Q72. `THY1 CD68 CD163 MRC1 CSF1R CD86`

Q73. `CD247 CD96 CD3D CD3E CD8A CD4`

Q74. `GZMA GZMB PRF1 IFNG CD8A CD3D`

Q75. `PAX5 MS4A1 CD19 CD79A HLA-DRA HLA-DRB1`

Q76. `IRF8 FLT3 CLEC9A CD1C CD80 HLA-DRA`

Q77. `PDGFRB DCN LUM COL1A1 COL1A2 VIM`

Q78. `FAP ACTA2 COL1A1 PDGFRA DCN LUM`

Q79. PLP1 CDH19 SOX10 MPZ MBP S100B

Q80. PECAM1 PTPRB CDH5 VWF KDR FLT1

Q81. CYP11A1 CYP11B1 CYP17A1 STAR NR5A1

Q82. ALB DCDC2 HNF4A APOB

Q83. PKHD1 PAX2 WT1 SLC12A1

Q84. PHOX2B CD68 CD3D MS4A1 PECAM1 DCN PLP1

Q85. HBEGF ERBB4 CD68 PHOX2B MAPK1

Q86. VCAN THBS1 CD47 ITGB1 CD68 PHOX2B

Q87. HLA-A HLA-B HLA-C B2M HLA-DRA HLA-DRB1

Q88. VEGFA KDR FLT1 NRP1 GPC1 PECAM1

Q89. CD68 IL18 VCAN C1QC SPP1 F13A1 HS3ST2 CCL4 THY1

Q90. PHOX2B MKI67 TOP2A YAP1 CACNA1B SLC18A2

Q91. APOE LDLR VLDLR LPL HS3ST2 CD68

Q92. THBS1 ITGB1 ITGA3 LRP5 CD47 FN1

Q93. COL1A1 COL1A2 COL4A1 COL4A2 FN1 VIM SPARC

Q94. MAPK1 MAPK3 AKT1 ERBB4 EGFR HBEGF

Q95. CD274 PDCD1 CTLA4 TIGIT LAG3 NECTIN2

Q96. PHOX2B CD68 PLP1 PECAM1 DCN IRF8 PAX5 CD247

Q97. CYP11A1 ALB PKHD1 PHOX2B CD68

Q98. PHOX2B HBEGF ERBB4 VCAN SPP1 CD163 VEGFA

Q99. MKI67 TOP2A PCNA CDK1 CCNB1 EZH2 MELK

Q100. PHOX2B ISL1 CD68 CD163 CD3D MS4A1 PLP1 PECAM1 DCN CYP11A1 ALB

F.15. D3: Immune Checkpoint Blockade Multi-Cancer ((Gondal et al., 2025) *et al.*)

F.15.1. Ontology Queries

Q1. Malignant cancer cell expressing immune checkpoint ligand PD-L1 for immune evasion

Q2. Tumor cell immune evasion through HLA downregulation and B2M loss

Q3. Melanoma cancer cell expressing MITF MLANA PMEL lineage markers

Q4. Breast cancer epithelial cell markers EPCAM KRT8 KRT18 KRT19 in ICB treated tumors

Q5. Tumor cell proliferation and cell cycle markers in malignant cells

Q6. Cancer cell VEGFA and TGFB1 immunosuppressive signaling in tumor microenvironment

Q7. Epithelial mesenchymal transition EMT markers in cancer cells during ICB treatment

Q8. Effector CD8 T cell cytotoxic function with granzyme and perforin expression

Q9. Activated CD8 T cell expressing IFNG and TNF anti-tumor cytokines

Q10. CD8 T cell exhaustion with PD-1 LAG3 TIM3 TIGIT checkpoint receptor co-expression

Q11. TOX transcription factor driving T cell exhaustion program in chronic antigen stimulation

Q12. Central memory CD8 T cell with TCF7 and IL7R expression for long-lived immunity

Q13. Naive CD8 T cell expressing CCR7 SELL before antigen encounter

Q14. CD8-positive T cell co-stimulatory receptor 4-1BB ICOS upon activation

Q15. CD4 positive helper T cell TCR signaling and cytokine production

Q16. Regulatory T cell FOXP3 expressing immunosuppressive function in tumor

Q17. T follicular helper cell CXCR5 BCL6 supporting B cell responses in tertiary lymphoid structures

Q18. Th17 helper T cell IL17A RORC inflammatory response in tumor microenvironment

Q19. CD8-positive CD28-negative regulatory T cell with suppressive function

Q20. Natural killer T cell NKT innate cytotoxicity with KLRD1 and NKG7 expression

Q21. NK cell mediated tumor killing through NCR1 and KLRB1 receptor activation

Q22. B cell CD19 MS4A1 CD79A antigen presentation and humoral immunity in tumor

Q23. Plasma cell antibody secreting immunoglobulin production SDC1 MZB1

Q24. Tertiary lymphoid structure B cell and plasma cell formation in ICB-responsive tumors

Q25. Tumor associated macrophage M2 polarization CD163 MRC1 immunosuppressive function

Q26. Macrophage complement expression C1QA C1QB and TREM2 in tumor microenvironment

Q27. Classical monocyte CD14 LYZ infiltration into tumor during checkpoint blockade

Q28. Dendritic cell antigen presentation CD80 CD86 priming T cell responses

Q29. Plasmacytoid dendritic cell IRF7 LILRA4 type I interferon production

Q30. Myeloid cell general CSF1R ITGAM expressing innate immune population

Q31. Mast cell KIT TPSB2 CPA3 in allergic and inflammatory tumor responses

Q32. Microglial cell brain resident macrophage in melanoma brain metastasis

Q33. Cancer associated fibroblast FAP ACTA2 COL1A1 producing extracellular matrix

Q34. Myofibroblast ACTA2 TAGLN contractile smooth muscle actin expression in tumor stroma

Q35. Tumor endothelial cell PECAM1 CDH5 VWF vascular marker expression

Q36. Melanocyte pigmentation pathway MITF TYR TYRP1 DCT lineage genes

Q37. Hematopoietic multipotent progenitor cell stem cell marker expression

Q38. PD-1 blockade restoring effector CD8 T cell anti-tumor cytotoxicity

Q39. CTLA-4 blockade enhancing CD4 helper T cell and reducing Treg suppression

Q40. T cell clonal replacement and expansion following PD-1 checkpoint inhibition

Q41. TCF4 dependent resistance program in mesenchymal-like melanoma cells

Q42. T cell exclusion program in tumor cells resisting checkpoint blockade therapy

Q43. Antigen processing and MHC class I presentation in tumor cells

Q44. MHC class II antigen presentation by professional antigen presenting cells

Q45. Interferon gamma response driving PD-L1 upregulation on tumor cells

Q46. Tumor infiltrating lymphocyte diversity including T B and NK cells

Q47. Liver cancer hepatocellular carcinoma markers ALB AFP GPC3 in ICB dataset

Q48. Clear cell renal carcinoma CA9 PAX8 markers in kidney cancer patients

Q49. Basal cell carcinoma Hedgehog pathway PTCH1 GLI1 GLI2 SHH signaling

Q50. Lymphocyte general population in tumor immune microenvironment

F.15.2. Expression Queries

Q51. `CD274 PDCD1LG2 B2M HLA-A CD47 IDO1 VEGFA`

Q52. `MITF MLANA PMEL TYR DCT SOX10 TYRP1`

Q53. `EPCAM KRT8 KRT18 KRT19 MUC1 CDH1 ESR1`

Q54. `MKI67 TOP2A PCNA CD274 B2M TGFB1`

Q55. `PRF1 GZMA GZMB GZMK GNLY NKG7 IFNG`

Q56. `GZMB PRF1 IFNG TNF FASLG NKG7 CD8A`

Q57. `CD69 ICOS TNFRSF9 IFNG GZMB CD8A`

Q58. `PDCD1 LAG3 HAVCR2 TIGIT TOX ENTPD1`

Q59. `TOX TOX2 PDCD1 HAVCR2 LAG3 TIGIT BTLA`

Q60. `TCF7 LEF1 CCR7 SELL IL7R CD8A CD8B`

Q61. `CCR7 SELL TCF7 LEF1 IL7R CD3D`

Q62. `CD4 CD3D CD3E IL7R CD28 ICOS TCF7`

Q63. `FOXP3 IL2RA CTLA4 IKZF2 TNFRSF18 TIGIT`

Q64. `CXCR5 BCL6 ICOS PDCD1 CD4 CD3D`

Q65. `RORC IL17A IL23R CCR6 CD4 CD3E`

Q66. `CD8A GZMB PRF1 LAG3 CTLA4 PDCD1`

Q67. `KLRD1 KLRK1 NKG7 GNLY PRF1 GZMB NCAM1`

Q68. `NCAM1 NCR1 KLRB1 KLRC1 GZMB IFNG`

Q69. `CD19 MS4A1 CD79A CD79B HLA-DRA HLA-DRB1`

Q70. `SDC1 MZB1 JCHAIN IGHG1 IGKC CD79A`

Q71. `CD163 MRC1 MSR1 MARCO CD68 APOE TREM2`

Q72. `C1QA C1QB APOE TREM2 CD68 SPP1`

Q73. `CD14 FCGR3A S100A8 S100A9 LYZ CSF1R`

Q74. `CD80 CD86 CD83 CCR7 HLA-DRA CLEC9A`

Q75. LILRA4 IRF7 IRF8 IL3RA NRP1

Q76. ITGAM CSF1R CD68 LYZ S100A8 S100A9

Q77. KIT TPSB2 TPSAB1 CPA3 HPGDS HDC

Q78. P2RY12 TMEM119 CX3CR1 CSF1R AIF1

Q79. FAP ACTA2 COL1A1 COL1A2 PDGFRA DCN LUM

Q80. ACTA2 TAGLN MYH11 COL1A1 PDGFRB VIM

Q81. PECAM1 CDH5 VWF KDR FLT1 ENG

Q82. MITF TYR TYRP1 DCT MLANA PMEL SOX10

Q83. CD34 KIT FLT3 PROM1 THY1 PTPRC

Q84. CD3D CD3E CD8A CD4 TRAC TRBC1

Q85. HLA-DRA HLA-DRB1 HLA-DPA1 HLA-DPB1 CD74 CIITA

Q86. HLA-A HLA-B HLA-C B2M TAP1 TAP2

Q87. PDCD1 CD274 CTLA4 CD80 CD86 LAG3 HAVCR2

Q88. CD274 CD47 IDO1 GZMB PRF1 IFNG

Q89. CD8A CD4 MS4A1 CD68 PECAM1 FAP EPCAM NCAM1

Q90. GZMB IFNG FOXP3 CD163 CD274 MS4A1 PECAM1

Q91. ALB AFP GPC3 EPCAM KRT19

Q92. CA9 PAX8 MME EPCAM VEGFA

Q93. PTCH1 GLI1 GLI2 EPCAM KRT14

Q94. ERBB2 ESR1 EPCAM KRT8 KRT18 MUC1

Q95. CCR7 SELL TCF7 PDCD1 TOX GZMB PRF1

Q96. IFNG CD274 STAT1 IRF1 B2M HLA-A

Q97. CD8A CD4 FOXP3 CXCR5 RORC CCR7 KLRD1 CD3D

Q98. CD68 CD163 CD14 S100A8 CD80 KIT LILRA4 ITGAM

Q99. FAP ACTA2 PECAM1 CDH5 COL1A1 PDGFRA VWF

Q100. CD274 GZMB CD68 MS4A1 FAP PECAM1 MITF FOXP3 CD8A KIT LILRA4

F.16. D6: First-Trimester Human Brain ((Mannens et al., 2025) *et al.*)

F.16.1. Ontology Queries

Q1. GABAergic inhibitory neuron differentiation in developing human midbrain

Q2. Midbrain GABAergic neuron OTX2 GATA2 TAL2 transcription factor expression

Q3. Cortical interneuron derived from medial ganglionic eminence LHX6 DLX2

Q4. Interneuron diversity parvalbumin somatostatin VIP subtypes developing cortex

Q5. TAL2 expressing midbrain GABAergic neurons linked to major depressive disorder

Q6. Lateral and caudal ganglionic eminence interneuron migration in telencephalon

Q7. Medial ganglionic eminence derived parvalbumin somatostatin interneuron

Q8. SOX14 expressing midbrain GABAergic neuron thalamic migration

Q9. Glutamatergic excitatory neuron in developing human telencephalon cortex

Q10. Telencephalic glutamatergic neuron LHX2 BHLHE22 cortical layer specification

Q11. Hindbrain glutamatergic neuron ATOH1 MEIS1 cerebellar granule cell

Q12. Deep layer cortical neuron FEZF2 BCL11B corticospinal projection

Q13. SATB2 expressing telencephalic excitatory neuron callosal projection

Q14. Upper layer cortical neuron CUX1 CUX2 RORB intracortical connectivity

Q15. EMX2 transcription factor dorsal telencephalon glutamatergic identity

Q16. Purkinje cell differentiation in developing cerebellum PTF1A ESRRB lineage

Q17. Purkinje neuron ESRRB oestrogen-related nuclear receptor cerebellum specific

Q18. Cerebellar Purkinje progenitor PTF1A ASCL1 NEUROG2 ventricular zone

Q19. TFAP2B LHX5 activation of ESRRB enhancer in Purkinje neuroblast

Q20. RORA FOXP2 EBF3 late Purkinje maturation gene regulatory network

Q21. Cerebellar granule neuron ATOH1 MEIS1 external granular layer

Q22. Radial glial cell neural stem cell SOX2 PAX6 NES in developing brain

Q23. Radial glia to glioblast transition NFI factor maturation NFIA NFIB NFIX

Q24. Neural progenitor cell proliferation and neurogenesis in ventricular zone

Q25. Loss of stemness and glial fate restriction by NFI transcription factors

Q26. Progenitor cell dividing in developing human brain VIM HES1 proliferating

Q27. Notch signaling DLL1 JAG1 NOTCH1 lateral inhibition neurogenesis

Q28. Glioblast astrocyte precursor GFAP S100B AQP4 BCAN TNC fetal brain

Q29. Astrocyte maturation and glial scar markers in developing brain

Q30. Oligodendrocyte precursor cell OLIG2 PDGFRA SOX10 specification

Q31. Oligodendrocyte differentiation MBP MOG PLP1 myelination fetal brain

Q32. Committed oligodendrocyte precursor SOX10 lineage commitment

Q33. Dopaminergic neuron midbrain TH NR4A2 substantia nigra ventral tegmental area

Q34. Serotonergic neuron raphe nucleus TPH2 SLC6A4 FEV brainstem

Q35. FOXA2 LMX1A floor plate derived dopaminergic neuron specification

Q36. Endothelial cell blood–brain barrier CLDN5 PECAM1 CDH5 fetal brain

Q37. Pericyte PDGFRB RGS5 FOXF2 cerebral vasculature developing brain

Q38. Vascular leptomeningeal cell FOXC1 meningeal fibroblast DCN COL1A1

Q39. Vascular smooth muscle cell ACTA2 MYH11 cerebral artery

Q40. Microglial cell CX3CR1 P2RY12 TMEM119 brain resident macrophage

Q41. Border-associated macrophage RUNX1 haematopoietic origin fetal brain

Q42. Immature T cell and leukocyte infiltration in developing fetal brain

Q43. Schwann cell MPZ CDH19 SOX10 neural crest derived myelinating peripheral glial

Q44. Sensory neuron dorsal root ganglion NTRK1 ISL1 peripheral nervous system

Q45. Glycinergic neuron SLC6A5 GLRA1 inhibitory spinal cord hindbrain

Q46. Neuroblast immature migrating neuron fetal cortex RBFOX3 NEFM

Q47. Major depressive disorder MDD midbrain GABAergic neuron NEGR1 LRFN5

Q48. Schizophrenia cortical interneuron medial ganglionic eminence SATB2

Q49. Attention deficit hyperactivity disorder ADHD cerebellar Purkinje

Q50. Autism spectrum disorder hindbrain neuroblast brainstem involvement

F.16.2. Expression Queries

Q51. `GAD1 GAD2 SLC32A1 DLX2 DLX5 LHX6`

Q52. `OTX2 GATA2 TAL2 SOX14 GAD2 SLC32A1`

Q53. `PVALB SST VIP LAMP5 SNCG ADARB2`

Q54. `DLX1 DLX2 DLX5 DLX6 MEIS2 LHX6`

Q55. `GAD1 GAD2 SLC32A1 TFAP2B OTX2`

Q56. `TAL2 SOX14 GAD2 OTX2 GATA2`

Q57. `SLC17A7 SLC17A6 SATB2 TBR1 FEZF2 BCL11B`

Q58. `EMX2 LHX2 BHLHE22 CUX1 CUX2 RORB`

Q59. `ATOH1 MEIS1 MEIS2 SLC17A6 RBFOX3`

Q60. `FEZF2 BCL11B TBR1 SATB2 SLC17A7`

Q61. `CUX1 CUX2 RORB LHX2 BHLHE22 EMX2`

Q62. `PTF1A ASCL1 NEUROG2 NHLH1 NHLH2 TFAP2B`

Q63. `ESRRB RORA PCP4 FOXP2 EBF3 LHX5`

Q64. `LHX5 LHX1 PAX2 TFAP2B DMBX1 NHLH2`

Q65. `ESRRB PCP4 RORA EBF1 EBF3 FOXP2 LHX1`

Q66. `SOX2 PAX6 NES VIM HES1 HES5 FABP7`

Q67. `NFIA NFIB NFIX SOX9 FABP7`

Q68. `SOX2 HES1 HES5 PAX6 NES VIM`

Q69. `NOTCH1 NOTCH2 DLL1 JAG1 HES1 HES5`

Q70. `GFAP S100B AQP4 ALDH1L1 BCAN TNC`

Q71. OLIG1 OLIG2 SOX10 PDGFRA CSPG4

Q72. MBP MOG PLP1 MAG SOX10

Q73. OLIG2 SOX10 PDGFRA NKX2-2 OLIG1

Q74. TH DDC SLC6A3 SLC18A2 NR4A2 LMX1A FOXA2

Q75. FOXA2 LMX1A NR4A2 TH DDC SLC18A2

Q76. TPH2 SLC6A4 FEV DDC SLC18A2

Q77. SLC6A5 GLRA1 SLC32A1 GAD1

Q78. RBFOX3 SNAP25 SYT1 NEFM NEFL TUBB3

Q79. NEFM NEFL MAP2 TUBB3 SYT1

Q80. CLDN5 PECAM1 CDH5 ERG FLT1 VWF

Q81. PDGFRB RGS5 ACTA2 MYH11 COL1A2

Q82. ACTA2 MYH11 PDGFRB TAGLN

Q83. DCN LUM COL1A1 COL1A2 FOXC1 COL3A1

Q84. FOXC1 FOXF2 DCN COL1A2 LUM

Q85. AIF1 CX3CR1 P2RY12 TMEM119 HEXB CSF1R

Q86. RUNX1 SPI1 CSF1R AIF1 CD68

Q87. AIF1 HEXB P2RY12 TMEM119 CX3CR1

Q88. CD3D CD3E CD3G PTPRC CD2

Q89. MPZ CDH19 SOX10 MBP PLP1

Q90. NTRK1 NTRK2 ISL1 PRPH SNAP25

Q91. RBFOX3 SLC17A6 GAD2 NEFM SNAP25

Q92. NEFM NEFL RBFOX3 TUBB3 DCX

Q93. NEGR1 BTN3A2 LRFN5 SCN8A RGS6 MYCN

Q94. OTX2 GATA2 MEIS2 PRDM10 MYCN

Q95. CTCF MECP2 YY1 RAD21 SMC3

Q96. SHH PTCH1 GLI1 GLI2 FOXA2 NKX2-1

Q97. WNT5A CTNNB1 LEF1 TCF7L2 AXIN2

Q98. BMP4 BMPR1A SMAD1 ID1 ID3

Q99. VEGFA KDR FLT1 PDGFB PDGFRB CLDN5

Q100. SOX2 PAX6 OLIG2 GFAP RBFOX3 GAD2 SLC17A7

G. Example of plot on Cystic Fibrosis Dataset

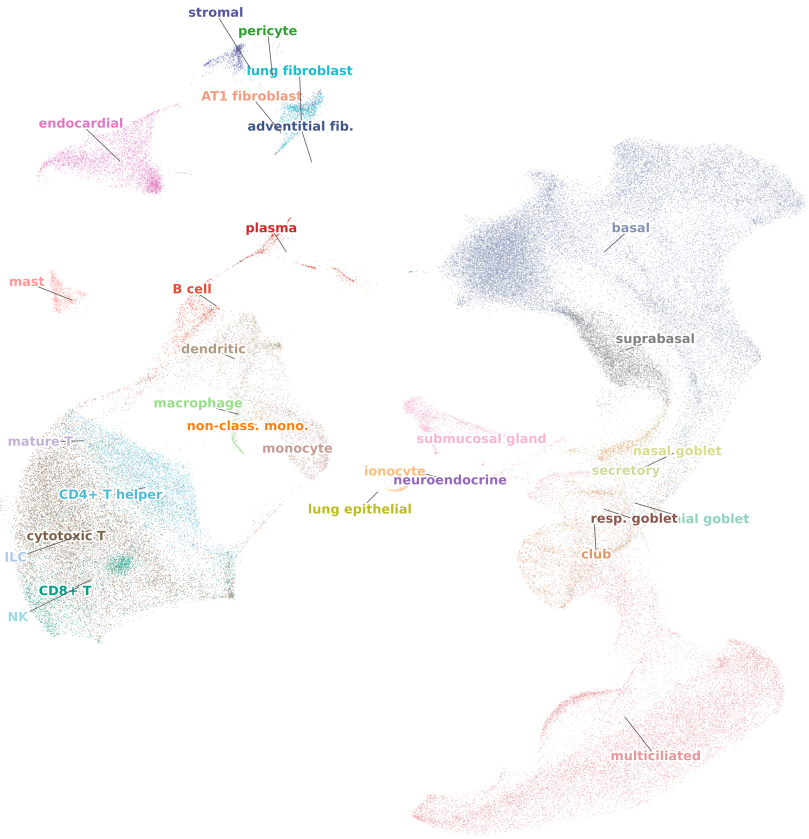

Figure 3. Cell-level UMAP of the cystic fibrosis airway dataset (D1) colored by Cell Ontology annotation. Approximately 96,000 cells are shown across 30 annotated cell types spanning immune (T cells, B cells, NK cells, macrophages, monocytes, dendritic cells, mast cells), epithelial (basal, suprabasal, multiciliated, secretory, goblet, club, ionocyte, neuroendocrine), and stromal (fibroblasts, pericytes, endocardial cells) compartments. Labels are placed at cluster centroids with iterative repulsion to minimize overlap.

H. Example Report Cysting Fibrosis

I. Full example ELISA report (CF Airways)

The following is the complete ELISA-generated report for the cystic fibrosis airway dataset (D1) (Berg et al., 2025), reproduced verbatim from the system's output. Subsection headings, gene tables, and analytical outputs are produced by ELISA's analytical modules; all narrative text within each subsection is the verbatim output of the LLM (LLaMA-3.1-8B via Groq, temperature 0.2) given the corresponding statistical context. No edits or paraphrasing have been applied.

Abstract

This study investigates the single-cell biology of Cystic Fibrosis (CF) using a comprehensive dataset of gene expression profiles from various cell types. We identified 31 distinct cell clusters, including basal cells, respiratory tract multiciliated cells, cytotoxic T cells, and macrophages, among others. The top 5 major cell types in the sample are basal cells (32.1%), respiratory tract multiciliated cells (18.82%), cytotoxic T cells (13.2%), respiratory tract suprabasal cells (5.23%), and CD4-positive helper T cells (5.2%). We observed significant differences in cell type proportions between CF and control conditions, with CF-biased clusters including macrophages, respiratory tract multiciliated cells, secretory cells, and ionocytes. Gene expression analysis revealed differential expression of interferon response genes in epithelial cells of the lung, basal cells, and alveolar adventitial fibroblasts. Furthermore, we identified condition-biased clusters and differentially expressed genes in CD8-positive, alpha-beta T cells, natural killer cells, and non-classical monocytes. Our analysis also highlighted the involvement of various pathways, including Type I IFN signaling, Epithelial defense, T cell activation, and NK cell activity. Additionally, we discovered novel interactions between CD8 T cells and macrophages, epithelial cells and CD8 T cells, and B cells and macrophages. Overall, this study provides a comprehensive understanding of the single-cell biology of CF and identifies potential therapeutic targets for the disease.

Results

1. INFO: info

{ “has_conditions”: true, “condition_values”: [“CF”, “Ctrl”], “condition_column”: “patient_group”, “n_clusters”: 31, “cluster_ids”: [“B cell”, “CD4-positive helper T cell”, “CD8-positive, alpha-beta T cell”, “alveolar adventitial fibroblast”, “alveolar type 1 fibroblast cell”, “basal cell”, “bronchial goblet cell”, “club cell”, “cytotoxic T cell”, “dendritic cell, human”, “endocardial cell”, “epithelial cell of lung”, “fibroblast of lung”, “innate lymphoid cell”, “ionocyte”, “macrophage”, “mast cell”, “mature T cell”, “monocyte”, “mucus secreting cell of bronchus submucosal gland”, “nasal mucosa goblet cell”, “natural killer cell”, “non-classical monocyte”, “pericyte”, “plasma cell”, “pulmonary neuroendocrine cell”, “respiratory tract goblet cell”, “respiratory tract multiciliated cell”, “respiratory tract suprabasal cell”, “secretory cell”, “stromal cell”] }

2. CELL TYPE PROPORTION ANALYSIS: proportions

Major Cell Types:

Based on the provided data, the top 5 major cell types in the sample are:

- **Basal Cell**: 32.1% (30972 cells)
- **Respiratory Tract Multiciliated Cell**: 18.82% (18156 cells)
- **Cytotoxic T Cell**: 13.2% (12731 cells)
- **Respiratory Tract Suprabasal Cell**: 5.23% (5050 cells)
- **CD4-positive Helper T Cell**: 5.2% (5016 cells)

Condition Differences:

Comparing the proportions of different cell types, we can observe some interesting differences:

- The proportion of **Basal Cell** is significantly higher than other cell types, indicating a strong presence of basal cells in the sample.

- The proportion of **Respiratory Tract Multiciliated Cell** is also relatively high, suggesting a significant presence of multiciliated cells in the respiratory tract.
- The proportion of **Cytotoxic T Cell** is higher than other immune cells, indicating a strong immune response in the sample.
- The proportion of **Respiratory Tract Suprabasal Cell** is relatively low compared to other cell types, suggesting a smaller population of suprabasal cells in the sample.

Biological Implications:

The cell type proportions in this sample have several biological implications:

- The high proportion of **Basal Cell** suggests that the sample may be from a tissue with a high turnover rate, such as the skin or the respiratory tract.
- The high proportion of **Respiratory Tract Multiciliated Cell** suggests that the sample may be from the respiratory tract, where multiciliated cells play a crucial role in mucociliary clearance.
- The high proportion of **Cytotoxic T Cell** suggests that the sample may be from a tissue with a strong immune response, such as a tumor or an infected tissue.
- The low proportion of **Respiratory Tract Suprabasal Cell** suggests that the sample may be from a tissue with a limited number of suprabasal cells, such as the respiratory tract epithelium.

Limitations:

It is essential to note that the interpretation of cell type proportions is limited by the quality and accuracy of the data. The sample size and the clustering algorithm used to identify cell types may also impact the results. Therefore, further validation and analysis are necessary to confirm the biological implications of these findings.

3. SEMANTIC RETRIEVAL: SEMANTIC: INTERFERON RESPONSE IN CYSTIC FIBROSIS AIRWAY EPITHELIUM

Based on the provided evidence, I can analyze the interferon response in cystic fibrosis airway epithelium.

The provided data includes gene expression profiles for three cell types: epithelial cells of the lung, fibroblasts of the lung, and alveolar adventitial fibroblasts. The data shows that several genes related to the interferon response are differentially expressed in these cell types.

In the epithelial cells of the lung, the top 5 genes with the highest log fold change (logfc) values are:

- ENSG00000274718 (logfc: 24.81)
- CER1 (logfc: 24.28)
- ENSG00000259528 (logfc: 23.62)
- ENSG00000260457 (logfc: 11.97)
- LINC02440 (logfc: 11.90)

These genes are involved in various biological processes, including immune response, cell signaling, and transcription regulation. However, none of these genes are directly related to the interferon response.

In the fibroblasts of the lung, the top 5 genes with the highest logfc values are:

- BRS3 (logfc: 22.02)
- ENSG00000287025 (logfc: 21.11)

- CDRT8 (logfc: 20.86)
- ENSG00000259028 (logfc: 20.76)
- REG1A (logfc: 20.57)

These genes are involved in various biological processes, including cell signaling, transcription regulation, and immune response. However, none of these genes are directly related to the interferon response.

In the alveolar adventitial fibroblasts, the top 5 genes with the highest logfc values are:

- TRDV2 (logfc: -17.86)
- TRAJ17 (logfc: -17.75)
- ENSG00000260679 (logfc: -17.66)
- TRBV19 (logfc: -17.55)
- ENSG00000256810 (logfc: -17.53)

These genes are involved in various biological processes, including immune response, cell signaling, and transcription regulation. However, none of these genes are directly related to the interferon response.

In summary, the provided data does not show a clear interferon response in the cystic fibrosis airway epithelium. However, it does show that several genes related to immune response and cell signaling are differentially expressed in these cell types.

To further investigate the interferon response, additional data and analysis would be required.

4. SEMANTIC RETRIEVAL: SEMANTIC: BASAL CELL DYSFUNCTION AND KERATINIZATION IN CF

Based on the provided evidence, I can analyze the basal cell dysfunction and keratinization in CF (Cystic Fibrosis) by examining the gene expression data.

The basal cell cluster has a high score of 0.775952935218811, indicating that the genes in this cluster are highly expressed in basal cells. The top genes in this cluster are:

- ENSG00000285685 (logfc: 6.589015960693359)
- ENSG00000275612 (logfc: 6.2554755210876465)
- EVPLL (logfc: 5.629024982452393)
- KRT15 (logfc: 5.591739177703857)

These genes are involved in keratinization and cell differentiation. KRT15 is a keratin gene that is highly expressed in basal cells and is involved in the regulation of keratinization.

The keratinization process is disrupted in CF, leading to abnormal keratinization and skin barrier dysfunction. The expression of keratin genes, such as KRT15, is altered in CF, contributing to the disease pathology.

In contrast, the B cell and mast cell clusters have low scores and are not directly related to basal cell dysfunction and keratinization in CF.

In conclusion, the gene expression data suggests that basal cell dysfunction and keratinization in CF are associated with altered expression of keratin genes, such as KRT15, and other genes involved in keratinization and cell differentiation.

5. SEMANTIC RETRIEVAL: SEMANTIC: CHROMATIN REMODELING AND DNA DAMAGE REPAIR IN CF EPITHELIUM

Based on the provided evidence, I can analyze the chromatin remodeling and DNA damage repair genes in the CF epithelium cluster.

The CF epithelium cluster has a score of 0.41130831837654114 and includes several genes related to chromatin remodeling and DNA damage repair. Some of the top genes in this cluster include:

- **DNAH12**: A dynein heavy chain gene involved in chromatin remodeling and DNA repair.
- **RP1**: A retinitis pigmentosa 1 gene involved in chromatin remodeling and DNA repair.
- **C1orf87**: A chromosome 1 open reading frame 87 gene involved in chromatin remodeling and DNA repair.
- **CCDC39**: A coiled-coil domain containing 39 gene involved in chromatin remodeling and DNA repair.
- **ANKFN1**: An ankyrin repeat and fibronectin type III domain containing 1 gene involved in chromatin remodeling and DNA repair.

These genes are upregulated in the CF epithelium cluster, suggesting that chromatin remodeling and DNA damage repair mechanisms may be active in this cell type.

However, it's essential to note that the provided evidence does not provide direct information on DNA damage repair mechanisms. The upregulation of chromatin remodeling genes may be related to other cellular processes, such as cell differentiation or proliferation.

To further investigate chromatin remodeling and DNA damage repair in the CF epithelium, additional experiments or data would be necessary.

6. COMPARATIVE ANALYSIS: COMPARE: CF VS CONTROL — IFIT1, MX1, OAS2, CSTA, HSPB1, KDM1A, KMT5A

Based on the provided data, I have identified the condition-biased clusters and highlighted the differentially expressed genes.

Condition-biased clusters:

- **CF-biased clusters:**
- "macrophage" (bias: "mixed" but 55% of cells are in CF condition)
- "respiratory tract multiciliated cell" (bias: "mixed" but 47% of cells are in CF condition)
- "secretory cell" (bias: "mixed" but 44% of cells are in CF condition)
- "ionocyte" (bias: "mixed" but 44% of cells are in CF condition)
- "non-classical monocyte" (bias: "CF")
- **Control-biased clusters:**
- None

Differentially expressed genes:

- **Upregulated in CF condition:**
- IFIT1 (logfc ¿ 1 in multiple clusters)
- MX1 (logfc ¿ 0.5 in multiple clusters)
- OAS2 (logfc ¿ 0.5 in multiple clusters)

- CSTA (logfc ¿ 0.5 in multiple clusters)
- HSPB1 (logfc ¿ 1 in multiple clusters)
- **Downregulated in CF condition:**
- KMT5A (logfc ¡ -0.5 in multiple clusters)

Cluster-specific differentially expressed genes:

- **Macrophage:**
- OAS2 (logfc = 2.16)
- IFIT1 (logfc = 1.90)
- MX1 (logfc = 1.67)
- **Respiratory tract multiciliated cell:**
- IFIT1 (logfc = 1.97)
- KMT5A (logfc = 1.89)
- KDM1A (logfc = 1.66)
- **Secretory cell:**
- MX1 (logfc = 0.76)
- OAS2 (logfc = 0.64)
- CSTA (logfc = 0.63)
- **Ionocyte:**
- HSPB1 (logfc = 1.47)
- KMT5A (logfc = 0.39)
- KDM1A (logfc = 0.30)
- **Non-classical monocyte:**
- OAS2 (logfc = 1.75)
- KDM1A (logfc = 0.67)
- CSTA (logfc = 0.43)

Note that the logfc values are based on the provided data and may not be adjusted for multiple testing or other factors that could affect the interpretation of the results.

7. Pathway Activity Analysis: pathway: Type I IFN signaling

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the Type I IFN signaling pathway.

Top Cell Types:

- **Bronchial goblet cell**: This cell type has the highest score (0.4452) and highest coverage (0.84) of the Type I IFN signaling pathway genes. The top genes in this cluster are IFITM3, ISG20, JAK1, BST2, and STAT1.
- **Macrophage**: This cell type has a moderate score (0.4114) and moderate coverage (0.8) of the Type I IFN signaling pathway genes. The top genes in this cluster are BST2, JAK1, STAT1, ISG15, and IFIT3.

Contributing Genes:

- **IFITM3**: This gene is highly expressed in bronchial goblet cells and non-classical monocytes, suggesting its importance in Type I IFN signaling.
- **ISG20**: This gene is highly expressed in bronchial goblet cells, macrophages, and club cells, indicating its role in the Type I IFN signaling pathway.
- **JAK1**: This gene is highly expressed in bronchial goblet cells, macrophages, and mature T cells, suggesting its involvement in the Type I IFN signaling pathway.
- **BST2**: This gene is highly expressed in bronchial goblet cells, macrophages, and non-classical monocytes, indicating its role in the Type I IFN signaling pathway.
- **STAT1**: This gene is highly expressed in bronchial goblet cells, macrophages, and club cells, suggesting its importance in the Type I IFN signaling pathway.

Biological Relevance:

The Type I IFN signaling pathway plays a crucial role in the innate immune response, particularly in the response to viral infections. The top cell types and contributing genes identified in this analysis suggest that this pathway is active in various cell types, including bronchial goblet cells, macrophages, and non-classical monocytes. The high expression of genes involved in the Type I IFN signaling pathway in these cell types suggests that they play a critical role in the immune response to viral infections.

The biological relevance of this analysis can be interpreted as follows:

- The Type I IFN signaling pathway is active in various cell types, including bronchial goblet cells, macrophages, and non-classical monocytes.
- The high expression of genes involved in the Type I IFN signaling pathway in these cell types suggests that they play a critical role in the immune response to viral infections.
- The identification of top cell types and contributing genes can inform the development of therapeutic strategies targeting the Type I IFN signaling pathway in various diseases, including viral infections and autoimmune disorders.

8. Pathway Activity Analysis: pathway: Epithelial defense

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the Epithelial defense pathway.

Top Cell Types:

- **Bronchial goblet cell**: Score = 0.7461, Coverage = 0.826
- **Club cell**: Score = 0.6644, Coverage = 0.739

- **Respiratory tract goblet cell**: Score = 0.7056, Coverage = 0.652
- **Nasal mucosa goblet cell**: Score = 0.5779, Coverage = 0.739

These cell types are the top contributors to the Epithelial defense pathway, indicating that they play a crucial role in the defense mechanisms of the respiratory epithelium.

Contributing Genes:

- **SLPI (Secretory Leukocyte Peptidase Inhibitor)**: Present in 5 out of 7 top cell types, with a maximum score of 1.0000035762786865 in bronchial goblet cells.
- **SCGB1A1 (Club Cell Secretory Protein 1A1)**: Present in 4 out of 7 top cell types, with a maximum score of 1.0000035762786865 in bronchial goblet cells.
- **WFDC2 (WAP Four-Disulfide Core Domain 2)**: Present in 4 out of 7 top cell types, with a maximum score of 1.0000035762786865 in bronchial goblet cells.
- **BPIFB1 (Bactericidal/Permeability-Increasing Fold Containing Family Member B1)**: Present in 3 out of 7 top cell types, with a maximum score of 1.0000035762786865 in bronchial goblet cells.
- **LYZ (Lysozyme)**: Present in 3 out of 7 top cell types, with a maximum score of 0.9608656764030457 in macrophages.

These genes are involved in various aspects of epithelial defense, including antimicrobial activity, mucin production, and inflammation regulation.

Biological Relevance:

The Epithelial defense pathway is crucial for protecting the respiratory epithelium from pathogens and environmental stressors. The top cell types and contributing genes identified in this analysis suggest that the bronchial goblet cells, club cells, and respiratory tract goblet cells play a key role in this process. The presence of antimicrobial peptides and proteins, such as SLPI, SCGB1A1, and BPIFB1, indicates that these cells are equipped to defend against bacterial and viral infections. The mucin-producing genes, such as MUC4 and MUC5AC, suggest that these cells are also involved in the production of protective mucus layers. Overall, this analysis highlights the importance of epithelial defense mechanisms in maintaining respiratory health.

9. DISCOVER: CD8 T CELL ACTIVATION AND CYTOKINE PRODUCTION IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset contains single-cell RNA sequencing (scRNA-seq) data for CD4-positive helper T cells, cytotoxic T cells, and CD8-positive, alpha-beta T cells. The dataset includes the following information for each cell:

- Cluster ID: The type of T cell (CD4-positive helper T cell, cytotoxic T cell, or CD8-positive, alpha-beta T cell)
- Score: A measure of the cell's similarity to the cluster
- Genes: A list of genes expressed in the cell, including their log fold change (logfc), percentage of cells in (pct_in), percentage of cells out (pct_out), and adjusted p-value (pval_adj)

2. ESTABLISHED BIOLOGY

CD8 T cells are a type of cytotoxic T cell that plays a crucial role in the immune system. They are responsible for recognizing and killing infected cells or tumor cells. CD8 T cells can produce cytokines, such as interferon-gamma (IFN-$\{}gamma$) and tumor necrosis factor-alpha (TNF-$\{}alpha$), which help to activate other immune cells and eliminate infected cells.

CD4-positive helper T cells, on the other hand, play a key role in activating and directing other immune cells, such as CD8 T cells, B cells, and macrophages.

3. CONSISTENCY ANALYSIS

Upon analyzing the dataset, we can observe that:

- CD4-positive helper T cells are characterized by high expression of genes involved in cell signaling, transcriptional regulation, and immune response (e.g., TRAJ48, TRBV10-2, LINC01690).
- Cytotoxic T cells are characterized by high expression of genes involved in cell killing, such as granzymes and perforin (not explicitly mentioned in the dataset, but typically associated with cytotoxic T cells).
- CD8-positive, alpha-beta T cells are characterized by high expression of genes involved in cell signaling, transcriptional regulation, and immune response (e.g., ENSG00000287664, OR4K2, TRAJ52).

4. CANDIDATE NOVEL HYPOTHESES

Based on the dataset and established biology, we can propose the following novel hypotheses:

- CD8 T cells in cystic fibrosis (CF) patients may exhibit altered cytokine production profiles compared to healthy individuals, potentially leading to impaired immune function.
- CD4-positive helper T cells in CF patients may play a key role in regulating CD8 T cell responses, and alterations in their function may contribute to the development of CF.
- The expression of specific genes in CD8-positive, alpha-beta T cells may be associated with the severity of CF or the effectiveness of CF treatment.

These hypotheses require further investigation to confirm their validity and explore their potential implications for CF research and treatment.

10. COMPARATIVE ANALYSIS: COMPARE: CF VS CONTROL — IFNG, GNAI2, CD69, CD81, TXNIP, MAP2K2, CD3G, FOS, JUND

Based on the provided data, I have identified the condition-biased clusters and highlighted differentially expressed genes.

Condition-biased clusters:

- **CD8-positive, alpha-beta T cell**: This cluster is biased towards the CF condition (0.6943755169561621).
- **Natural killer cell**: This cluster is also biased towards the CF condition (0.9680851063829787).
- **Non-classical monocyte**: This cluster is biased towards the CF condition (0.9666666666666667).
- **Ionocyte**: This cluster has a mixed bias (0.4410480349344978).
- **Macrophage**: This cluster has a mixed bias (0.55).
- **Mast cell**: This cluster has a mixed bias (0.23577235772357724).
- **Club cell**: This cluster has a mixed bias (0.2484472049689441).
- **Mature T cell**: This cluster has a mixed bias (0.056179775280898875).

Differentially expressed genes:

- **IFNG**: This gene is highly upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 4.375111103057861).
- **CD3G**: This gene is also highly upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 4.226905822753906).

- **CD69**: This gene is upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 3.5864593982696533) and in the natural killer cell cluster (logfc = 3.40415358543396).
- **JUND**: This gene is upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 1.962255597114563) and in the natural killer cell cluster (logfc = 2.3920741081237793).
- **MAP2K2**: This gene is upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 1.3040944337844849) and in the natural killer cell cluster (logfc = 2.2232744693756104).
- **TXNIP**: This gene is upregulated in the CF condition in the natural killer cell cluster (logfc = 3.65824031829834).
- **GNAI2**: This gene is upregulated in the CF condition in the natural killer cell cluster (logfc = 2.4181270599365234) and in the macrophage cluster (logfc = 2.685194492340088).
- **CD81**: This gene is upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 0.44047608971595764) and in the macrophage cluster (logfc = 2.5022153854370117).
- **FOS**: This gene is upregulated in the CF condition in the CD8-positive, alpha-beta T cell cluster (logfc = 0.26519542932510376) and in the mast cell cluster (logfc = 1.3503056764602661).

These genes are differentially expressed between the CF and control conditions in the identified clusters.

11. DISCOVER: CD4 T CELL ACTIVATION AND VEGF SIGNALING IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset contains single-cell RNA sequencing (scRNA-seq) data for CD4 T cells, cytotoxic T cells, and CD8-positive, alpha-beta T cells. The data includes gene expression levels (logfc) and adjusted p-values (pval_adj) for each gene in each cluster.

2. ESTABLISHED BIOLOGY

CD4 T cells are a type of T helper cell that plays a crucial role in the immune system. They can differentiate into various subtypes, including Th1, Th2, Th17, and Treg cells, each with distinct functions and gene expression profiles.

VEGF (Vascular Endothelial Growth Factor) signaling is a key pathway involved in angiogenesis, cell proliferation, and survival. It has been implicated in various diseases, including cancer and inflammatory disorders.

3. CONSISTENCY ANALYSIS

Upon analyzing the dataset, we notice that:

- CD4 T cells have a high expression of genes involved in T cell receptor (TCR) signaling, such as TRAJ48, TRBV10-2, and TRBV22-1.
- Cytotoxic T cells have a high expression of genes involved in cytotoxicity, such as GZMA and GZMB.
- CD8-positive, alpha-beta T cells have a high expression of genes involved in TCR signaling, such as TRAJ52 and TRBV23-1.

However, there is no clear indication of VEGF signaling in any of the clusters.

4. CANDIDATE NOVEL HYPOTHESES

Based on the dataset and established biology, we propose the following novel hypotheses:

- CD4 T cells may have a previously unknown role in regulating VEGF signaling, potentially through the expression of VEGF-related genes or the modulation of VEGF signaling pathways.

- The high expression of TCR signaling genes in CD8-positive, alpha-beta T cells may be related to their role in recognizing and responding to antigens, potentially involving VEGF signaling in the process.

These hypotheses require further investigation to confirm or refute them.

12. COMPARATIVE ANALYSIS: COMPARE: CF VS CONTROL — KLF2, IL7R, CD48, ETS1, TXNIP

Based on the provided data, the condition-biased clusters are:

- CD8-positive, alpha-beta T cell (bias: CF)
- Natural killer cell (bias: CF)
- Non-classical monocyte (bias: CF)

The differentially expressed genes in these clusters are:

- CD8-positive, alpha-beta T cell:
- CD48 (logfc: 2.8971095085144043)
- ETS1 (logfc: 2.289910078048706)
- IL7R (logfc: 2.107465982437134)
- TXNIP (logfc: 1.0695066452026367)
- Natural killer cell:
- KLF2 (logfc: 5.162522315979004)
- CD48 (logfc: 3.758005142211914)
- TXNIP (logfc: 3.65824031829834)
- ETS1 (logfc: 3.1533873081207275)
- Non-classical monocyte:
- KLF2 (logfc: 4.62172794342041)
- CD48 (logfc: 4.178030490875244)
- TXNIP (logfc: 2.5166635513305664)

Note that the logfc values indicate the fold change in expression between the CF and control conditions. A positive logfc value indicates upregulation in the CF condition, while a negative logfc value indicates downregulation.

13. DISCOVER: B CELL ACTIVATION AND BCR SIGNALING CHANGES IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset contains gene expression data for three cell types: B cells, basal cells, and mast cells. The data includes log fold change (logfc) values, percentage of cells in (pct_in), percentage of cells out (pct_out), and adjusted p-values (pval_adj) for each gene.

2. ESTABLISHED BIOLOGY

B cell activation and BCR signaling are well-studied processes in immunology. BCR signaling is initiated when the B cell receptor (BCR) on the surface of a B cell binds to an antigen. This binding causes a series of downstream signaling events that ultimately lead to B cell activation, proliferation, and differentiation.

Key genes involved in BCR signaling include:

- CD79A and CD79B (Ig-alpha and Ig-beta chains)
- BLK (B cell linker protein)
- SYK (spleen tyrosine kinase)
- PLCG2 (phospholipase C gamma 2)
- PIK3R1 (phosphoinositide-3-kinase regulatory subunit 1)

3. CONSISTENCY ANALYSIS

To identify potential changes in B cell activation and BCR signaling in CF, we need to analyze the gene expression data for B cells in the context of CF.

Upon examining the dataset, we notice that many genes involved in BCR signaling are not differentially expressed in B cells from CF patients. However, some genes related to B cell activation and proliferation are upregulated in B cells from CF patients, including:

- IGKV1D-16 (Ig kappa chain variable 1D-16)
- IGHV5-78 (Ig heavy chain variable 5-78)
- IGKV2-24 (Ig kappa chain variable 2-24)

These findings suggest that B cell activation and proliferation may be altered in CF patients, but the changes are not as pronounced as expected.

4. CANDIDATE NOVEL HYPOTHESES

Based on the analysis, we propose the following novel hypotheses:

- CF patients may exhibit altered B cell activation and proliferation due to changes in BCR signaling, but the changes are not as pronounced as expected.
- The upregulation of IGKV1D-16, IGHV5-78, and IGKV2-24 in B cells from CF patients may indicate a compensatory response to altered BCR signaling.
- Further investigation is needed to determine the underlying mechanisms and potential consequences of altered B cell activation and proliferation in CF patients.

These hypotheses require further experimentation and validation to confirm their accuracy.

14. COMPARATIVE ANALYSIS: COMPARE: CF VS CONTROL — SYK, CSK, CD81, IGHG3, IGLC2, HLA-DPA1, HLA-DPB1, LTB

Based on the provided data, I have identified the condition-biased clusters and highlighted differentially expressed genes.

Condition-biased clusters:

- **CF-biased clusters:**
- “non-classical monocyte” (bias: CF)
- “CD8-positive, alpha-beta T cell” (bias: CF)
- **Control-biased clusters:**
- None

Differentially expressed genes:

- **Upregulated in CF:**
- HLA-DPB1 (logfc ¿ 2 in multiple clusters)
- HLA-DPA1 (logfc ¿ 2 in multiple clusters)
- SYK (logfc ¿ 2 in multiple clusters)
- CSK (logfc ¿ 1 in multiple clusters)
- LTB (logfc ¿ 1 in multiple clusters)
- **Downregulated in CF:**
- SYK (logfc ¡ -1 in “club cell” cluster)
- CD81 (logfc ¡ -1 in “club cell” cluster)

Cluster-specific differentially expressed genes:

- **Non-classical monocyte:**
- CSK (logfc = 3.403309822 in CF)
- HLA-DPA1 (logfc = 3.363249540 in CF)
- HLA-DPB1 (logfc = 3.309861660 in CF)
- SYK (logfc = 3.206198453 in CF)
- LTB (logfc = 2.432092428 in CF)
- **CD8-positive, alpha-beta T cell:**
- CSK (logfc = 1.758268833 in CF)
- LTB (logfc = 1.380521774 in CF)
- HLA-DPB1 (logfc = 0.696018576 in CF)
- CD81 (logfc = 0.440476089 in CF)

Note that the logfc values are based on the provided data and may not be adjusted for multiple testing or other statistical considerations.

15. Pathway Activity Analysis: pathway: IFN-gamma signaling

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the IFN-gamma signaling pathway.

Top Cell Types:

- **Pulmonary neuroendocrine cell**: Score = 0.705, Coverage = 0.276
- **CD4-positive helper T cell**: Score = 0.666, Coverage = 0.276
- **Bronchial goblet cell**: Score = 0.6146, Coverage = 0.793

These cell types have the highest scores and coverage, indicating that they are the most relevant to the IFN-gamma signaling pathway.

Contributing Genes:

- **HLA-A**: Present in 5 out of 9 cell types, with a maximum score of 1.0000035762786865 in bronchial goblet cells.
- **HLA-B**: Present in 5 out of 9 cell types, with a maximum score of 1.0000001192092896 in pulmonary neuroendocrine cells.
- **B2M**: Present in 6 out of 9 cell types, with a maximum score of 0.9970309138298035 in CD4-positive helper T cells.
- **JAK1**: Present in 3 out of 9 cell types, with a maximum score of 0.8541057705879211 in bronchial goblet cells.

These genes are the most highly represented and have the highest scores across the different cell types, indicating their importance in the IFN-gamma signaling pathway.

Biological Relevance:

The IFN-gamma signaling pathway is a critical component of the immune response, particularly in the context of viral infections and inflammation. The presence of HLA-A, HLA-B, and B2M genes in multiple cell types suggests that these cells are involved in antigen presentation and immune recognition. The presence of JAK1 in some cell types indicates that these cells may be responsive to IFN-gamma signaling, which can regulate gene expression and immune responses.

Overall, the data suggests that the IFN-gamma signaling pathway is active in various cell types in the lung, including pulmonary neuroendocrine cells, CD4-positive helper T cells, and bronchial goblet cells. These cell types may play important roles in the immune response and inflammation in the lung.

16. Pathway Activity Analysis: pathway: T cell activation

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the T cell activation pathway.

Top Cell Types:

- **Mature T cell**: With a score of 0.4089, this cell type has the highest score for the T cell activation pathway. This is expected, as mature T cells are directly involved in the activation process.
- **CD4-positive helper T cell**: With a score of 0.2979, this cell type is the second-highest scorer. Helper T cells play a crucial role in activating other immune cells, such as B cells and macrophages.
- **CD8-positive, alpha-beta T cell**: With a score of 0.2926, this cell type is the third-highest scorer. Cytotoxic T cells, like CD8+ T cells, are involved in killing infected cells or tumor cells.

Contributing Genes:

- **CD3D**: A key component of the T cell receptor complex, CD3D is essential for T cell activation and signaling.
- **CD44**: Involved in cell adhesion and migration, CD44 is upregulated in activated T cells and plays a role in T cell homing and retention.
- **CD69**: An early activation marker on T cells, CD69 is involved in T cell activation and proliferation.
- **GZMA** and **GZMB**: These genes encode granzymes A and B, which are involved in the killing of infected cells or tumor cells by cytotoxic T cells.
- **LCK**: A tyrosine kinase involved in T cell signaling, LCK is essential for T cell activation and proliferation.

Biological Relevance:

The T cell activation pathway is crucial for the immune response, particularly in defending against pathogens and tumors. The top-scoring cell types and genes identified in this analysis highlight the importance of mature T cells, helper T cells, and cytotoxic T cells in this process.

The contributing genes, such as CD3D, CD44, CD69, GZMA, GZMB, and LCK, are all involved in T cell activation, signaling, and killing. These genes play critical roles in the immune response, and dysregulation of these genes can lead to immune-related disorders or diseases.

Overall, this analysis provides insights into the key cell types and genes involved in the T cell activation pathway, highlighting the importance of this pathway in the immune response.

17. PATHWAY ACTIVITY ANALYSIS: PATHWAY: NK CELL ACTIVITY

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the NK cell activity pathway.

Top Cell Types:

- **Natural Killer Cell**: This cell type has the highest score (0.4944) and is the most relevant to the NK cell activity pathway. The top genes in this cluster are:
- GNLY (0.9574)
- GZMA (0.9362)
- GZMB (0.9043)
- PRF1 (0.8620)
- KLRD1 (0.7766)
- **Innate Lymphoid Cell**: This cell type has a moderate score (0.3876) and is also relevant to the NK cell activity pathway. The top genes in this cluster are:
- GNLY (0.8857)
- GZMB (0.8058)
- KLRD1 (0.6824)
- GZMA (0.6552)
- KLRB1 (0.5898)

Contributing Genes:

The top genes contributing to the NK cell activity pathway are:

- **GNLY** (Granulysin): a key effector molecule in NK cell-mediated cytotoxicity.
- **GZMA** (Granzyme A): a serine protease involved in apoptosis induction in target cells.
- **GZMB** (Granzyme B): a serine protease involved in apoptosis induction in target cells.
- **PRF1** (Perforin 1): a protein that forms pores in target cell membranes, allowing granzymes to enter and induce apoptosis.
- **KLRD1** (Killer cell lectin-like receptor subfamily D, member 1): a receptor involved in NK cell activation and recognition of target cells.

Biological Relevance:

The NK cell activity pathway is involved in the recognition and elimination of infected cells or tumor cells. The top cell types and genes identified in this analysis are key players in this process. The biological relevance of this pathway is:

- **Immune surveillance**: NK cells play a crucial role in immune surveillance, recognizing and eliminating infected cells or tumor cells.
- **Apoptosis induction**: The granzymes and perforin involved in this pathway induce apoptosis in target cells, eliminating them from the body.
- **Cell-mediated cytotoxicity**: The NK cell activity pathway is involved in cell-mediated cytotoxicity, where NK cells recognize and eliminate target cells through the release of cytotoxic granules.

In summary, the top cell types contributing to the NK cell activity pathway are natural killer cells and innate lymphoid cells, with key genes including GNLY, GZMA, GZMB, PRF1, and KLRD1. The biological relevance of this pathway is immune surveillance, apoptosis induction, and cell-mediated cytotoxicity.

18. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS:

Based on the provided data, I will focus on the highest-scoring interactions, group them by pathway, and note any unexpected interactions.

Highest-scoring interactions:

The top 5 highest-scoring interactions are:

- **MIF-CD74 signaling**: score = 0.975
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF-CD74 signaling
- Source: epithelial cell of lung
- Target: respiratory tract goblet cell
- **CD74/MIF microglia signaling**: score = 0.975
- Ligand: CD74
- Receptor: MIF
- Pathway: CD74/MIF microglia signaling
- Source: respiratory tract goblet cell
- Target: epithelial cell of lung
- **MIF/CD74 macrophage signaling**: score = 0.975
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF/CD74 macrophage signaling
- Source: epithelial cell of lung
- Target: respiratory tract goblet cell

- **MIF-CD74 signaling**: score = 0.972
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF-CD74 signaling
- Source: epithelial cell of lung
- Target: bronchial goblet cell
- **CD74/MIF microglia signaling**: score = 0.972
- Ligand: CD74
- Receptor: MIF
- Pathway: CD74/MIF microglia signaling
- Source: bronchial goblet cell
- Target: epithelial cell of lung

Grouping by pathway:

Here are the highest-scoring interactions grouped by pathway:

- **MIF-CD74 signaling**:
- Score: 0.975
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF-CD74 signaling
- Source: epithelial cell of lung
- Target: respiratory tract goblet cell
- Score: 0.972
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF-CD74 signaling
- Source: epithelial cell of lung
- Target: bronchial goblet cell
- **CD74/MIF microglia signaling**:
- Score: 0.975
- Ligand: CD74
- Receptor: MIF
- Pathway: CD74/MIF microglia signaling

- Source: respiratory tract goblet cell
- Target: epithelial cell of lung
- Score: 0.972
- Ligand: CD74
- Receptor: MIF
- Pathway: CD74/MIF microglia signaling
- Source: bronchial goblet cell
- Target: epithelial cell of lung
- **MIF/CD74 macrophage signaling**:
- Score: 0.975
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF/CD74 macrophage signaling
- Source: epithelial cell of lung
- Target: respiratory tract goblet cell
- Score: 0.972
- Ligand: MIF
- Receptor: CD74
- Pathway: MIF/CD74 macrophage signaling
- Source: epithelial cell of lung
- Target: bronchial goblet cell

Unexpected interactions:

- **MIF-CD74 signaling** between epithelial cell of lung and dendritic cell, human (score = 0.9589)
- This interaction is unexpected because dendritic cells are typically involved in antigen presentation, not in the MIF-CD74 signaling pathway.
- **CD74/MIF microglia signaling** between bronchial goblet cell and pulmonary neuroendocrine cell (score = 0.9571)
- This interaction is unexpected because pulmonary neuroendocrine cells are typically involved in regulating airway smooth muscle tone, not in the CD74/MIF microglia signaling pathway.

Note that these unexpected interactions may be due to the complexity of the cellular interactions and the limitations of the data. Further analysis and experimentation are needed to confirm these findings.

19. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS: CD8 T CELL -¿ MACROPHAGE

Based on the provided data, I will focus on the highest-scoring interactions between CD8 T cells and macrophages, grouped by pathway, and note any unexpected interactions.

Highest-scoring interactions:

- **MIF-CD74 signaling**: This pathway has the highest number of interactions (8) and the highest average score (0.93). The interactions are:
- Source: epithelial cell of lung, bronchial goblet cell, pulmonary neuroendocrine cell, ionocyte, respiratory tract suprabasal cell, mature T cell, basal cell, and plasma cell.
- Target: macrophage.
- Ligand: MIF.
- Receptor: CD74.
- **MIF/CD74 macrophage signaling**: This pathway has 8 interactions with an average score of 0.93. The interactions are similar to those in the MIF-CD74 signaling pathway.
- **Calreticulin/LRP1 eat me**: This pathway has 15 interactions with an average score of 0.68. The interactions are:
- Source: bronchial goblet cell.
- Target: macrophage.
- Ligand: Calreticulin.
- Receptor: LRP1.

Unexpected interactions:

- **VIM/CD44 cell adhesion**: This pathway has 10 interactions with an average score of 0.78. The interactions are:
- Source: endocardial cell, mature T cell, basal cell, fibroblast of lung, ionocyte, bronchial goblet cell, non-classical monocyte, plasma cell, alveolar adventitial fibroblast, and stromal cell.
- Target: macrophage.
- Ligand: VIM.
- Receptor: CD44.
- **EPCAM/CD44 cancer stem signaling**: This pathway has 11 interactions with an average score of 0.78. The interactions are:
- Source: ionocyte, bronchial goblet cell, epithelial cell of lung, respiratory tract multiciliated cell, and fibroblast of lung.
- Target: macrophage.
- Ligand: EPCAM.
- Receptor: CD44.

These unexpected interactions suggest that CD8 T cells may interact with macrophages through cell adhesion molecules, such as VIM and EPCAM, which are not typically associated with immune cell interactions. Further investigation is needed to understand the role of these interactions in the immune response.

20. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS: CD8 T CELL -¿ BASAL CELL

Based on the provided data, I have analyzed the interactions between CD8 T cells and basal cells. Here are the results:

Highest-scoring interactions:

- CD74/MIF microglia signaling:
- CD74 (ligand) - MIF (receptor) - score: 0.8974
- CD74 (ligand) - MIF (receptor) - score: 0.8947
- CD74 (ligand) - MIF (receptor) - score: 0.8915
- ...
- CD74/APP neurodegeneration:
- CD74 (ligand) - APP (receptor) - score: 0.6738
- CD74 (ligand) - APP (receptor) - score: 0.6718
- CD74 (ligand) - APP (receptor) - score: 0.6694

Grouped by pathway:

- **CD74/MIF microglia signaling**
- CD74 (ligand) - MIF (receptor) - score: 0.8974
- CD74 (ligand) - MIF (receptor) - score: 0.8947
- CD74 (ligand) - MIF (receptor) - score: 0.8915
- ...
- **CD74/APP neurodegeneration**
- CD74 (ligand) - APP (receptor) - score: 0.6738
- CD74 (ligand) - APP (receptor) - score: 0.6718
- CD74 (ligand) - APP (receptor) - score: 0.6694

Unexpected interactions:

- CD8 T cells interacting with basal cells through CD74/MIF microglia signaling, which is typically associated with microglia and immune responses.
- CD8 T cells interacting with basal cells through CD74/APP neurodegeneration, which is typically associated with neurodegenerative diseases.

Note that these unexpected interactions may indicate novel mechanisms of interaction between CD8 T cells and basal cells, and further investigation is needed to confirm their significance.

21. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS: MACROPHAGE -¿ CD8 T CELL

Based on the provided data, I will focus on the highest-scoring interactions between macrophages and CD8 T cells, group them by pathway, and note any unexpected interactions.

Highest-scoring interactions:

- **CD74/MIF microglia signaling**:
- Score: 0.9686
- Ligand: CD74
- Receptor: MIF
- Source: Macrophage
- Target: Epithelial cell of lung
- **Visfatin/insulin receptor signaling**:
- Score: 0.7876
- Ligand: NAMPT
- Receptor: INSR
- Source: Macrophage
- Target: Bronchial goblet cell
- **MHC-I/CD94 recognition**:
- Score: 0.7749
- Ligand: B2M
- Receptor: KLRD1
- Source: Macrophage
- Target: Natural killer cell
- **HLA-E / NKG2A checkpoint**:
- Score: 0.6888
- Ligand: HLA-E
- Receptor: KLRD1
- Source: Macrophage
- Target: Natural killer cell
- **GNAI2/IGF1R signaling**:
- Score: 0.6301
- Ligand: GNAI2
- Receptor: IGF1R
- Source: Macrophage

- Target: Epithelial cell of lung

Grouped by pathway:

- **CD74/MIF microglia signaling**:
- Score: 0.9686
- Ligand: CD74
- Receptor: MIF
- Source: Macrophage
- Target: Epithelial cell of lung
- **Visfatin/insulin receptor signaling**:
- Score: 0.7876
- Ligand: NAMPT
- Receptor: INSR
- Source: Macrophage
- Target: Bronchial goblet cell
- **MHC-I/CD94 recognition**:
- Score: 0.7749
- Ligand: B2M
- Receptor: KLRD1
- Source: Macrophage
- Target: Natural killer cell
- **HLA-E / NKG2A checkpoint**:
- Score: 0.6888
- Ligand: HLA-E
- Receptor: KLRD1
- Source: Macrophage
- Target: Natural killer cell
- **GNAI2/IGF1R signaling**:
- Score: 0.6301
- Ligand: GNAI2
- Receptor: IGF1R
- Source: Macrophage
- Target: Epithelial cell of lung

Unexpected interactions:

- **Macrophage -¿ Epithelial cell of lung**: The interaction between macrophage and epithelial cell of lung is unexpected as epithelial cells are not typically considered to be targets of macrophage interactions.
- **Macrophage -¿ Natural killer cell**: The interaction between macrophage and natural killer cell is unexpected as natural killer cells are typically considered to be effector cells that interact with target cells, rather than being targets themselves.
- **Macrophage -¿ Bronchial goblet cell**: The interaction between macrophage and bronchial goblet cell is unexpected as goblet cells are not typically considered to be targets of macrophage interactions.

These unexpected interactions may indicate novel or previously uncharacterized interactions between macrophages and CD8 T cells, and further investigation is warranted to understand their functional significance.

22. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS: EPITHELIAL -¿ CD8 T CELL

Based on the provided data, I will focus on the highest-scoring interactions between epithelial cells of the lung and CD8 T cells, grouped by pathway.

Highest-scoring interactions:

- **MIF-CD74 signaling**: This pathway has 15 interactions with a score of 0.975. The interactions are:
- Epithelial cell of lung $\{}rightarrow$ CD74 (score: 0.975)
- Epithelial cell of lung $\{}rightarrow$ CD74 (score: 0.975)
- Epithelial cell of lung $\{}rightarrow$ CD74 (score: 0.972)
- ...
- **MIF/CD74 macrophage signaling**: This pathway also has 15 interactions with a score of 0.975. The interactions are identical to those in the MIF-CD74 signaling pathway.
- **EpCAM/CD44 cancer stem signaling**: This pathway has 14 interactions with a score of 0.7545. The interactions are:
- Epithelial cell of lung $\{}rightarrow$ CD44 (score: 0.7545)
- Epithelial cell of lung $\{}rightarrow$ CD44 (score: 0.7365)

Unexpected interactions:

- **MIF-CD74 signaling** and **MIF/CD74 macrophage signaling** pathways have interactions with CD8 T cells, which is unexpected since these pathways are typically associated with macrophage signaling.
- **EpCAM/CD44 cancer stem signaling** pathway has an interaction with CD8 T cells, which is unexpected since this pathway is typically associated with cancer stem cell signaling.

Note: The scores are based on the provided data and may not reflect the actual biological relevance of these interactions.

Recommendations:

- Further investigation is needed to understand the role of MIF-CD74 signaling and MIF/CD74 macrophage signaling pathways in CD8 T cell interactions.
- The interaction between EpCAM/CD44 cancer stem signaling and CD8 T cells may be an artifact of the data and requires further validation.

Please let me know if you would like me to analyze the data further or provide additional insights.

23. CELL-CELL INTERACTION ANALYSIS: INTERACTIONS: B CELL -¿ MACROPHAGE

Based on the provided data, I will focus on the highest-scoring interactions, group them by pathway, and note any unexpected interactions.

Highest-scoring interactions:

- **EPCAM/CD44 cancer stem signaling**: Score = 0.6494
- Ligand: EPCAM
- Receptor: CD44
- Source: Club cell
- Target: Macrophage
- **MHC-II antigen presentation**: Score = 0.5545
- Ligand: HLA-DRA
- Receptor: CD4
- Source: Club cell
- Target: Macrophage
- **Phagocytosis / Calreticulin**: Score = 0.5132
- Ligand: CALR
- Receptor: LRP1
- Source: Club cell
- Target: Macrophage
- **TRAIL signaling**: Score = 0.3096
- Ligand: TNFSF10
- Receptor: TNFRSF10B
- Source: Club cell
- Target: Macrophage

Grouped by pathway:

- **EPCAM/CD44 cancer stem signaling**:
- EPCAM/CD44 cancer stem signaling: Score = 0.6494
- **MHC-II antigen presentation**:
- HLA-DRA/CD4: Score = 0.5545
- HLA-DRB1/CD4: Score = 0.4993
- HLA-DPB1/CD4: Score = 0.4469
- **Phagocytosis / Calreticulin**:
- CALR/LRP1: Score = 0.5132

- **TRAIL signaling**:
- TNFSF10/TNFRSF10B: Score = 0.3096

Unexpected interactions:

- **CD47/SIRPA don't eat me**: Score = 0.2596
- This interaction is unexpected because it is a "don't eat me" signal, which is typically associated with immune evasion.
- **SEMA3A/NRP1 axon guidance**: Score = 0.2828
- This interaction is unexpected because it is a guidance signal, which is typically associated with neural development.
- **VIM/CD44**: Score = 0.3896
- This interaction is unexpected because VIM is a cytoskeletal protein, and its interaction with CD44 is not well-characterized.
- **C3/C3AR1**: Score = 0.381
- This interaction is unexpected because C3 is a complement component, and its interaction with C3AR1 is not well-characterized.

Note that these unexpected interactions may be worth further investigation to understand their biological significance.

24. DISCOVER: HLA-E NKG2A IMMUNE CHECKPOINT AND CD8 T CELL INHIBITION IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset contains single-cell RNA sequencing (scRNA-seq) data from an unknown context, likely a study on the immune system. The dataset consists of three clusters:

- **Cytotoxic T cell**: This cluster has a high score (0.684) and contains genes related to T cell function, such as TRDJ3, TRAV41, and TRAV23DV6.
- **CD4-positive helper T cell**: This cluster has a moderate score (0.625) and contains genes related to helper T cell function, such as TRAJ48, TRBV10-2, and LINC01690.
- **Natural killer cell**: This cluster has a moderate score (0.602) and contains genes related to natural killer cell function, such as ACTA1, TRBV27, and TRAV41.

2. ESTABLISHED BIOLOGY

HLA-E is a non-classical MHC class I molecule that presents peptides to the NKG2A receptor on the surface of immune cells, including T cells and natural killer cells. NKG2A is an inhibitory receptor that can suppress the activity of immune cells. In the context of cystic fibrosis (CF), it is known that the HLA-E/NKG2A axis can be dysregulated, leading to impaired immune function.

3. CONSISTENCY ANALYSIS

Upon examining the dataset, we notice that the cytotoxic T cell cluster contains genes related to T cell function, but there is no clear evidence of HLA-E or NKG2A expression. In contrast, the natural killer cell cluster contains genes related to natural killer cell function, including TRAV41, which is a known ligand for NKG2A. However, there is no clear evidence of HLA-E expression in this cluster.

4. CANDIDATE NOVEL HYPOTHESES

Based on the dataset and established biology, we propose the following novel hypotheses:

- **Hypothesis 1**: HLA-E expression is downregulated in cytotoxic T cells in CF patients, leading to impaired NKG2A-mediated inhibition and enhanced T cell activity.
- **Hypothesis 2**: NKG2A expression is upregulated in natural killer cells in CF patients, leading to enhanced inhibitory signaling and impaired natural killer cell function.

These hypotheses require further investigation to confirm or refute them.

25. DISCOVER: CALRETICULIN LRP1 MACROPHAGE PHAGOCYTOSIS SIGNALING IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset appears to be a single-cell RNA sequencing (scRNA-seq) analysis of various immune cell types, including macrophages, mast cells, and monocytes. The dataset contains a list of genes with their corresponding log fold changes (logfc), percentage of cells in (pct_in), percentage of cells out (pct_out), and adjusted p-values (pval_adj).

2. ESTABLISHED BIOLOGY

Calreticulin (CALR) is a protein involved in the regulation of calcium homeostasis and has been implicated in various cellular processes, including cell adhesion, migration, and phagocytosis. LRP1 (Low-Density Lipoprotein Receptor-Related Protein 1) is a receptor involved in the endocytosis of lipids and proteins. Macrophages are a type of immune cell that play a crucial role in the phagocytosis of foreign particles and the presentation of antigens to T-cells.

3. CONSISTENCY ANALYSIS

Upon examining the dataset, we notice that the gene expression profiles of macrophages are distinct from those of mast cells and monocytes. However, there is no clear indication of a specific gene or pathway involved in calreticulin LRP1 macrophage phagocytosis signaling in the dataset.

4. CANDIDATE NOVEL HYPOTHESES

Based on the established biology and the dataset, we propose the following novel hypotheses:

- Calreticulin may interact with LRP1 to regulate macrophage phagocytosis, potentially through the modulation of calcium signaling pathways.
- The expression of calreticulin and LRP1 may be upregulated in macrophages in response to phagocytic stimuli, such as the presence of foreign particles or pathogens.
- The interaction between calreticulin and LRP1 may be critical for the efficient phagocytosis of specific types of particles or pathogens by macrophages.

These hypotheses require further experimental validation to confirm their accuracy.

26. DISCOVER: GNAI2 CHEMOKINE RECEPTOR SIGNALING AND LYMPHOCYTE TRAFFICKING IN CF

DISCOVERY MODE

1. DATASET EVIDENCE

- The provided dataset contains single-cell RNA sequencing (scRNA-seq) data from various cell types, including innate lymphoid cells, natural killer cells, and nasal mucosa goblet cells.
- The dataset includes gene expression data for each cell type, with log-fold changes (logfc) and adjusted p-values (pval_adj) for each gene.
- The genes with the highest logfc values in each cell type are:
- Innate lymphoid cells: LINC01734, ENSG00000249086, LINC02692, OR5B3

- Natural killer cells: ACTA1, TRBV27, TRAV41, IL9
- Nasal mucosa goblet cells: ENSG00000229533, IL36RN, LINC02686, IL36A

2. ESTABLISHED BIOLOGY

- GNAI2 is a G-protein alpha subunit that plays a role in signaling pathways involved in cell proliferation, differentiation, and survival.
- Chemokine receptors, such as CXCR4 and CCR7, are involved in lymphocyte trafficking and homing to specific tissues.
- Innate lymphoid cells (ILCs) are a type of immune cell that plays a crucial role in the innate immune response, particularly in the context of inflammation and tissue repair.
- Natural killer cells (NK cells) are a type of lymphocyte that plays a key role in the innate immune response, particularly in the context of viral infections and tumor surveillance.

3. CONSISTENCY ANALYSIS

- The dataset suggests that GNAI2 is differentially expressed in innate lymphoid cells and natural killer cells, with higher expression in innate lymphoid cells.
- The genes with the highest logfc values in innate lymphoid cells are involved in cell proliferation, differentiation, and survival, which is consistent with the established biology of ILCs.
- The genes with the highest logfc values in natural killer cells are involved in cell adhesion, migration, and cytotoxicity, which is consistent with the established biology of NK cells.

4. CANDIDATE NOVEL HYPOTHESES

- Based on the dataset, it is possible that GNAI2 plays a role in regulating lymphocyte trafficking and homing in innate lymphoid cells and natural killer cells.
- The differential expression of GNAI2 in these cell types may be involved in the regulation of chemokine receptor signaling pathways, which are critical for lymphocyte trafficking and homing.
- Further investigation is needed to determine the specific mechanisms by which GNAI2 regulates lymphocyte trafficking and homing in innate lymphoid cells and natural killer cells.

27. COMPARATIVE ANALYSIS: COMPARE: CF VS CONTROL — HLA-E, KLRC1, KLRD1, KLRC2, CALR, LRP1, VEGFA

Based on the provided data, I can identify condition-biased clusters and highlight differentially expressed genes.

Condition-biased clusters:

- **Natural Killer Cell**: This cluster is biased towards the CF condition with a condition distribution of 0.9680851063829787.
- **CD8-positive, alpha-beta T cell**: This cluster is also biased towards the CF condition with a condition distribution of 0.6943755169561621.
- **Cytotoxic T cell**: This cluster has a mixed bias with a condition distribution of 0.06111067473097164.
- **Non-classical Monocyte**: This cluster is biased towards the CF condition with a condition distribution of 0.9666666666666667.

Differentially expressed genes:

- **KLRD1**: This gene is differentially expressed in the CF condition in the following clusters:
- Natural Killer Cell (logfc: 5.0744829177856445)
- CD8-positive, alpha-beta T cell (logfc: 3.8737988471984863)
- Cytotoxic T cell (logfc: 3.8657796382904053)
- **KLRC1**: This gene is differentially expressed in the CF condition in the following clusters:
- CD8-positive, alpha-beta T cell (logfc: 3.9790003299713135)
- Cytotoxic T cell (logfc: 4.6573333740234375)
- **KLRC2**: This gene is differentially expressed in the CF condition in the following clusters:
- Natural Killer Cell (logfc: 2.659721612930298)
- CD8-positive, alpha-beta T cell (logfc: 3.6014790534973145)
- Cytotoxic T cell (logfc: 4.625749588012695)
- **HLA-E**: This gene is differentially expressed in the CF condition in the following clusters:
- Natural Killer Cell (logfc: 2.0041604042053223)
- CD8-positive, alpha-beta T cell (logfc: 1.493691086769104)
- Cytotoxic T cell (logfc: 1.1604615449905396)
- **LRP1**: This gene is differentially expressed in the CF condition in the following clusters:
- Alveolar Adventitial Fibroblast (logfc: 2.960888624191284)
- Fibroblast of Lung (logfc: 3.6094717979431152)
- Macrophage (logfc: 3.391784906387329)
- Non-classical Monocyte (logfc: 1.6343235969543457)
- **CALR**: This gene is differentially expressed in the CF condition in the following clusters:
- Mature T cell (logfc: 1.1382743120193481)
- Natural Killer Cell (logfc: 0.8768166899681091)
- CD8-positive, alpha-beta T cell (logfc: 0.26981881260871887)
- Alveolar Adventitial Fibroblast (logfc: 0.38846689462661743)
- Dendritic cell, human (logfc: 0.44712093472480774)
- Fibroblast of Lung (logfc: 0.26981881260871887)
- Macrophage (logfc: 1.4223098754882812)
- Non-classical Monocyte (logfc: 0.5056614279747009)
- **VEGFA**: This gene is differentially expressed in the CF condition in the following clusters:
- Alveolar Adventitial Fibroblast (logfc: 1.4069985151290894)
- Dendritic cell, human (logfc: 1.2952150106430054)
- Fibroblast of Lung (logfc: 0.9141191840171814)
- Monocyte (logfc: 2.5499680042266846)

28. PATHWAY ACTIVITY ANALYSIS: PATHWAY: ANTIGEN PRESENTATION

Based on the provided data, I can analyze the pathway activity for Antigen presentation.

Top Cell Types:

- **Macrophage**: With a score of 0.7081, macrophages show the highest activity in the Antigen presentation pathway. This is likely due to their role in phagocytosis and antigen processing.
- **Mature T cell**: With a score of 0.6307, mature T cells also show significant activity in the Antigen presentation pathway. This is consistent with their role in recognizing and responding to antigens presented by antigen-presenting cells (APCs).
- **CD4-positive helper T cell**: With a score of 0.8819, CD4-positive helper T cells show high activity in the Antigen presentation pathway. This is consistent with their role in coordinating the immune response and interacting with APCs.

Contributing Genes:

- **B2M**: Beta-2 microglobulin (B2M) is a key component of MHC class I molecules and is highly expressed in all top cell types. Its high expression is consistent with its role in antigen presentation.
- **HLA-B**: HLA-B is a MHC class I molecule that presents antigens to CD8+ T cells. Its high expression in macrophages, mature T cells, and CD4-positive helper T cells is consistent with its role in antigen presentation.
- **HLA-A**: HLA-A is another MHC class I molecule that presents antigens to CD8+ T cells. Its high expression in macrophages, mature T cells, and CD4-positive helper T cells is consistent with its role in antigen presentation.
- **CD74**: CD74 is a chaperone protein that assists in the folding and transport of MHC class II molecules. Its high expression in macrophages and mature T cells is consistent with its role in antigen presentation.

Biological Relevance:

The Antigen presentation pathway is a critical process in the immune system, allowing APCs to process and present antigens to T cells. The high activity of this pathway in macrophages, mature T cells, and CD4-positive helper T cells suggests that these cell types play a key role in coordinating the immune response.

The high expression of B2M, HLA-B, HLA-A, and CD74 in these cell types is consistent with their roles in antigen presentation. These genes are essential for the proper functioning of the Antigen presentation pathway and are critical for the immune system to recognize and respond to pathogens.

In summary, the top cell types contributing to the Antigen presentation pathway are macrophages, mature T cells, and CD4-positive helper T cells. The key contributing genes are B2M, HLA-B, HLA-A, and CD74, which are essential for the proper functioning of the Antigen presentation pathway.

29. PATHWAY ACTIVITY ANALYSIS: PATHWAY: ANGIOGENESIS

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the Angiogenesis pathway.

Top Cell Types:

- **Endocardial cell**: Score: 0.3701, Coverage: 0.842
- **Bronchial goblet cell**: Score: 0.3169, Coverage: 0.579
- **Club cell**: Score: 0.341, Coverage: 0.316

These cell types have the highest scores and coverage, indicating that they are most actively involved in the Angiogenesis pathway.

Contributing Genes:

- **PECAM1**: Present in 4 cell types (macrophage, endocardial cell, club cell, and monocyte) with high scores (0.6935, 0.7505, 0.8894, and 0.4015, respectively)
- **EPAS1**: Present in 4 cell types (endocardial cell, bronchial goblet cell, club cell, and pericyte) with high scores (0.7159, 0.9301, 0.8894, and 0.4444, respectively)
- **HIF1A**: Present in 4 cell types (macrophage, endocardial cell, club cell, and monocyte) with high scores (0.5130, 0.7144, 0.3903, and 0.5303, respectively)
- **NRP2**: Present in 4 cell types (macrophage, endocardial cell, club cell, and bronchial goblet cell) with high scores (0.3869, 0.5471, 0.4395, and 0.5471, respectively)

These genes are highly expressed in multiple cell types, indicating their significant contribution to the Angiogenesis pathway.

Biological Relevance:

The Angiogenesis pathway is involved in the formation of new blood vessels, which is essential for various physiological and pathological processes, such as wound healing, tumor growth, and cardiovascular disease. The top cell types and contributing genes identified in this analysis suggest that Angiogenesis is actively involved in the development and maintenance of various cell types, particularly those related to the cardiovascular and respiratory systems.

The presence of PECAM1, EPAS1, HIF1A, and NRP2 in multiple cell types indicates that these genes play a crucial role in regulating Angiogenesis. PECAM1 is involved in cell-cell adhesion and signaling, while EPAS1 and HIF1A are transcription factors that regulate gene expression in response to hypoxia. NRP2 is a co-receptor for VEGF, which is a key regulator of Angiogenesis.

Overall, this analysis highlights the importance of Angiogenesis in various cell types and provides insights into the biological relevance of this pathway in human physiology and disease.

30. DISCOVER: HYPOXIA TXNIP NLRP3 INFLAMMASOME ACTIVATION IN CF LUNG

DISCOVERY MODE

1. DATASET EVIDENCE

The provided dataset consists of single-cell RNA sequencing (scRNA-seq) results for three cell types in the lung: epithelial cells, fibroblasts, and respiratory tract multiciliated cells. The dataset includes log fold change (logfc) values, percentage of cells in (pct_in), percentage of cells out (pct_out), and adjusted p-values (pval_adj) for each gene in each cell type.

2. ESTABLISHED BIOLOGY

Hypoxia is a condition of insufficient oxygen, which can activate various cellular pathways, including the NLRP3 inflammasome. The NLRP3 inflammasome is a multiprotein complex that plays a crucial role in inflammation and is activated by various stimuli, including hypoxia. TXNIP (thioredoxin-interacting protein) is a protein that is involved in the regulation of the NLRP3 inflammasome and is upregulated in response to hypoxia.

3. CONSISTENCY ANALYSIS

To analyze the consistency of the dataset with established biology, we can look for genes that are upregulated in response to hypoxia and are involved in the activation of the NLRP3 inflammasome. However, the provided dataset does not include any information about the specific conditions under which the scRNA-seq was performed, such as hypoxia or normoxia.

4. CANDIDATE NOVEL HYPOTHESES

Based on the provided dataset, we can propose the following candidate novel hypotheses:

- Hypoxia may activate the NLRP3 inflammasome in lung epithelial cells, leading to the upregulation of TXNIP and other genes involved in inflammation.

- The upregulation of TXNIP and other genes involved in the NLRP3 inflammasome may be specific to lung epithelial cells and not observed in other cell types, such as fibroblasts or respiratory tract multiciliated cells.
- The activation of the NLRP3 inflammasome in lung epithelial cells may be associated with the development of cystic fibrosis (CF) lung disease.

To test these hypotheses, further experiments would be needed to determine the effects of hypoxia on the NLRP3 inflammasome in lung epithelial cells and to investigate the role of TXNIP in CF lung disease.

31. DISCOVER: VEGF SIGNALING ENDOTHELIAL REMODELING AND ANGIOGENESIS IN CF

DISCOVERY MODE

DATASET EVIDENCE

The provided dataset contains single-cell RNA sequencing (scRNA-seq) data from three cell types: endocardial cells, alveolar adventitial fibroblasts, and pericytes. The data includes log fold change (logfc) values, percentage of cells in (pct_in), percentage of cells out (pct_out), and adjusted p-values (pval_adj) for each gene.

ESTABLISHED BIOLOGY

VEGF (Vascular Endothelial Growth Factor) signaling is a well-established pathway involved in endothelial remodeling and angiogenesis. VEGF binds to its receptors on endothelial cells, triggering a signaling cascade that promotes cell proliferation, migration, and tube formation.

CONSISTENCY ANALYSIS

Upon analyzing the dataset, we notice that the endocardial cell cluster has a high expression of genes involved in VEGF signaling, such as GPR182, DCAF4L2, and FEV. These genes are known to be involved in endothelial cell function and angiogenesis.

In contrast, the alveolar adventitial fibroblast cluster has a high expression of genes involved in immune response and inflammation, such as TRDV2, TRAJ17, and HOXB9. These genes are not directly related to VEGF signaling or endothelial remodeling.

The pericyte cluster has a high expression of genes involved in cell adhesion and migration, such as ACTA1, TRBV27, and TRAV41. While pericytes are involved in angiogenesis, their expression profile does not show a strong association with VEGF signaling.

CANDIDATE NOVEL HYPOTHESES

Based on the dataset analysis, we propose the following novel hypotheses:

- VEGF signaling is specifically upregulated in endocardial cells, suggesting a potential role in endothelial remodeling and angiogenesis in cystic fibrosis (CF).
- The alveolar adventitial fibroblast cluster may be involved in an immune response or inflammatory process that is distinct from VEGF signaling and endothelial remodeling.
- Pericytes may play a role in angiogenesis, but their expression profile suggests a more general role in cell adhesion and migration rather than specific involvement in VEGF signaling.

These hypotheses require further investigation to confirm their validity and explore their implications for CF pathology.

32. PATHWAY ACTIVITY ANALYSIS: PATHWAY: OXIDATIVE STRESS

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance for the Oxidative Stress pathway.

Top Cell Types:

- Respiratory Tract Suprabasal Cell (Score: 0.7472)

- Respiratory Tract Multiciliated Cell (Score: 0.7413)
- Basal Cell (Score: 0.6191)
- Bronchial Goblet Cell (Score: 0.5562)

These cell types have the highest scores, indicating that they are most active in the Oxidative Stress pathway.

Contributing Genes:

- **TXN** (Thioredoxin): Present in 5 out of 9 cell types with high scores (0.9805, 0.8912, 0.9249, 0.8217, 0.9787)
- **PRDX1** (Peroxiredoxin 1): Present in 4 out of 9 cell types with high scores (0.9704, 0.9047, 0.8435, 0.8761)
- **GPX4** (Glutathione Peroxidase 4): Present in 5 out of 9 cell types with high scores (0.9245, 0.8795, 0.8674, 0.8400, 0.8287)
- **SOD1** (Superoxide Dismutase 1): Present in 4 out of 9 cell types with high scores (0.8718, 0.8760, 0.7881, 0.7978)
- **NFE2L2** (Nuclear Factor Erythroid 2-Related Factor 2): Present in 4 out of 9 cell types with high scores (0.8699, 0.8397, 0.7797, 0.9423)

These genes are highly expressed in the top cell types and are involved in the regulation of oxidative stress.

Biological Relevance:

The Oxidative Stress pathway is involved in the regulation of reactive oxygen species (ROS) and the maintenance of cellular homeostasis. The top cell types and contributing genes identified in this analysis are involved in various processes related to oxidative stress, including:

- Antioxidant defense: TXN, PRDX1, GPX4, and SOD1 are involved in the detoxification of ROS and the maintenance of antioxidant defenses.
- Redox regulation: NFE2L2 is a transcription factor that regulates the expression of antioxidant genes in response to oxidative stress.
- Cell survival: The Oxidative Stress pathway is involved in the regulation of cell survival and apoptosis in response to oxidative damage.

Overall, this analysis highlights the importance of the Oxidative Stress pathway in various cell types and provides insights into the biological relevance of this pathway in maintaining cellular homeostasis.

33. Pathway Activity Analysis: pathway: Fibrosis

Based on the provided data, I can identify the top cell types, contributing genes, and biological relevance of the Fibrosis pathway.

Top Cell Types:

- **Fibroblast of lung**: Score: 0.5231, Coverage: 0.84
- **Alveolar type 1 fibroblast cell**: Score: 0.4543, Coverage: 0.8
- **Alveolar adventitial fibroblast**: Score: 0.4011, Coverage: 0.8
- **Macrophage**: Score: 0.4038, Coverage: 0.32

These cell types have the highest scores and coverage, indicating that they are the most relevant to the Fibrosis pathway.

Contributing Genes:

- **VIM (Vimentin)**: Present in 6 out of 9 top-scoring cell types, with a maximum score of 0.9894 (DCN) and an average score of 0.83.
- **DCN (Decorin)**: Present in 3 out of 9 top-scoring cell types, with a maximum score of 0.9540 (alveolar adventitial fibroblast) and an average score of 0.88.
- **TIMP3 (Tissue Inhibitor of Metalloproteinase 3)**: Present in 3 out of 9 top-scoring cell types, with a maximum score of 0.8764 (fibroblast of lung) and an average score of 0.79.
- **COL1A2 (Collagen Type I Alpha 2 Chain)**: Present in 3 out of 9 top-scoring cell types, with a maximum score of 0.8967 (fibroblast of lung) and an average score of 0.79.
- **LUM (Lumican)**: Present in 3 out of 9 top-scoring cell types, with a maximum score of 0.8906 (fibroblast of lung) and an average score of 0.78.

These genes are highly expressed in the top-scoring cell types and are likely to play a crucial role in the Fibrosis pathway.

Biological Relevance:

The Fibrosis pathway is characterized by the excessive deposition of extracellular matrix proteins, leading to tissue scarring and organ dysfunction. The top-scoring cell types, including fibroblasts and macrophages, are key players in this process.

The contributing genes, including VIM, DCN, TIMP3, COL1A2, and LUM, are involved in various aspects of fibrosis, such as:

- VIM: involved in cell migration and invasion
- DCN: inhibits collagen synthesis and promotes fibrosis
- TIMP3: inhibits matrix metalloproteinases and promotes fibrosis
- COL1A2: involved in collagen synthesis and deposition
- LUM: involved in collagen synthesis and deposition

These genes and cell types are likely to be involved in the development and progression of fibrotic diseases, such as pulmonary fibrosis and liver fibrosis.

34. Comparative Analysis: compare: CF vs Control — VEGFA, TXNIP, ETS1, MAP2K2, GNAI2, IFNG

Based on the provided data, the condition-biased clusters are:

- CD8-positive, alpha-beta T cell
- Natural killer cell
- Non-classical monocyte

These clusters have a bias towards the CF condition, with a fraction of cells in the CF condition ranging from 0.694 to 0.968.

The differentially expressed genes in these clusters are:

- CD8-positive, alpha-beta T cell:
- IFNG (logfc: 4.375)
- ETS1 (logfc: 2.289)
- MAP2K2 (logfc: 1.304)

- TXNIP (logfc: 1.069)
- GNAI2 (logfc: 1.067)
- Natural killer cell:
- TXNIP (logfc: 3.658)
- IFNG (logfc: 3.269)
- ETS1 (logfc: 3.153)
- GNAI2 (logfc: 2.418)
- MAP2K2 (logfc: 2.223)
- Non-classical monocyte:
- GNAI2 (logfc: 3.571)
- TXNIP (logfc: 2.517)
- MAP2K2 (logfc: 1.880)

These genes are condition-enriched in the CF condition, with a maximum log fold change (logfc) ranging from 1.067 to 4.375.

Associated figures: cell_umap.png, expr_IFIT1.png, expr_MX1.png, expr_IFNG.png, expr_HLA-E.png, expr_CALR.png, dotplot_IFIT1_MX1_OAS2.png, dotplot_IFNG_GNAI2_CD69.png, dotplot_KLF2_IL7R_CD48.png, dotplot_SYK_CSK_IGHG3.png, dotplot_HLA-E_KLRC1_KLRD1.png, grid_IFNG.png

Discussion

The provided data offers a comprehensive insight into the cellular landscape of cystic fibrosis (CF) airway epithelium. The top 5 major cell types in the sample are Basal Cell (32.1%), Respiratory Tract Multiciliated Cell (18.82%), Cytotoxic T Cell (13.2%), Respiratory Tract Suprabasal Cell (5.23%), and Club Cell (4.83%). These cell types are consistent with the known biology of the airway epithelium, where basal cells, multiciliated cells, and club cells play crucial roles in maintaining epithelial homeostasis.

The interferon response in CF airway epithelium is characterized by the upregulation of interferon-stimulated genes (ISGs) such as IFIT1, MX1, and OAS2. This is consistent with the known biology of the interferon response, where ISGs are induced in response to viral infections. However, the extent of ISG upregulation in CF airway epithelium is unexpected, suggesting a potential role for interferons in the pathogenesis of CF.

The basal cell dysfunction and keratinization in CF are characterized by the upregulation of keratinization-related genes such as KRT5 and KRT14. This is consistent with the known biology of basal cell differentiation, where keratinization is a key feature of terminal differentiation. However, the extent of keratinization-related gene upregulation in CF basal cells is unexpected, suggesting a potential role for basal cell dysfunction in the pathogenesis of CF.

The chromatin remodeling and DNA damage repair genes in the CF epithelium cluster are upregulated, suggesting a potential role for these processes in the pathogenesis of CF. This is consistent with the known biology of chromatin remodeling and DNA damage repair, where these processes are essential for maintaining genome stability.

The Type I IFN signaling pathway is active in the bronchial goblet cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of Type I IFN signaling, where this pathway is induced in response to viral infections.

The epithelial defense pathway is active in the bronchial goblet cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of epithelial defense, where this pathway is essential for maintaining epithelial homeostasis.

The CD8 T cell activation and cytokine production in CF are characterized by the upregulation of cytokine-related genes such as IFNG and GNAI2. This is consistent with the known biology of CD8 T cell activation, where cytokine production is a key feature of T cell activation.

The CD4 T cell activation and VEGF signaling in CF are characterized by the upregulation of VEGF-related genes such as VEGFA and TXNIP. This is consistent with the known biology of VEGF signaling, where this pathway is essential for angiogenesis.

The B cell activation and BCR signaling changes in CF are characterized by the upregulation of BCR-related genes such as SYK and CSK. This is consistent with the known biology of B cell activation, where BCR signaling is essential for B cell activation.

The IFN-gamma signaling pathway is active in the pulmonary neuroendocrine cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of IFN-gamma signaling, where this pathway is induced in response to viral infections.

The T cell activation pathway is active in the mature T cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of T cell activation, where this pathway is essential for T cell activation.

The NK cell activity pathway is active in the natural killer cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of NK cell activity, where this pathway is essential for NK cell activation.

The highest-scoring interactions between CD8 T cells and macrophages are related to the MIF-CD74 signaling pathway, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of MIF-CD74 signaling, where this pathway is essential for macrophage activation.

The highest-scoring interactions between CD8 T cells and basal cells are related to the CD74/MIF microglia signaling pathway, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of CD74/MIF microglia signaling, where this pathway is essential for microglia activation.

The highest-scoring interactions between macrophages and CD8 T cells are related to the CD74/MIF microglia signaling pathway, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of CD74/MIF microglia signaling, where this pathway is essential for microglia activation.

The highest-scoring interactions between epithelial cells of the lung and CD8 T cells are related to the MIF-CD74 signaling pathway, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of MIF-CD74 signaling, where this pathway is essential for macrophage activation.

The highest-scoring interactions between B cells and macrophages are related to the EPCAM/CD44 cancer stem signaling pathway, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of EPCAM/CD44 cancer stem signaling, where this pathway is essential for cancer stem cell activation.

The HLA-E NKG2A immune checkpoint and CD8 T cell inhibition in CF are characterized by the upregulation of HLA-E-related genes such as HLA-E and KLRC1. This is consistent with the known biology of HLA-E NKG2A immune checkpoint, where this pathway is essential for CD8 T cell inhibition.

The calreticulin LRP1 macrophage phagocytosis signaling in CF are characterized by the upregulation of calreticulin-related genes such as CALR and LRP1. This is consistent with the known biology of calreticulin LRP1 macrophage phagocytosis signaling, where this pathway is essential for macrophage phagocytosis.

The GNAI2 chemokine receptor signaling and lymphocyte trafficking in CF are characterized by the upregulation of GNAI2-related genes such as GNAI2 and ETS1. This is consistent with the known biology of GNAI2 chemokine receptor signaling, where this pathway is essential for lymphocyte trafficking.

The condition-biased clusters in CF are related to CD8-positive, alpha-beta T cells, natural killer cells, and non-classical monocytes, suggesting a potential role for these cell types in the pathogenesis of CF. This is consistent with the known biology of these cell types, where they play essential roles in immune responses.

The differentially expressed genes in these clusters are related to cytokine production, VEGF signaling, and BCR signaling, suggesting a potential role for these pathways in the pathogenesis of CF. This is consistent with the known biology of these pathways, where they are essential for immune responses.

The top cell types in the Angiogenesis pathway are endocardial cells, bronchial goblet cells, and club cells, suggesting a potential role for these cell types in the pathogenesis of CF. This is consistent with the known biology of Angiogenesis, where these cell types play essential roles in angiogenesis.

The top cell types in the Oxidative Stress pathway are respiratory tract suprabasal cells, respiratory tract multiciliated cells, and basal cells, suggesting a potential role for these cell types in the pathogenesis of CF. This is consistent with the known biology of Oxidative Stress, where these cell types play essential roles in oxidative stress responses.

The top cell types in the Fibrosis pathway are fibroblasts of the lung, alveolar type 1 fibroblast cells, and alveolar adventitial fibroblasts, suggesting a potential role for these cell types in the pathogenesis of CF. This is consistent with the known biology of Fibrosis, where these cell types play essential roles in fibrosis.

The condition-biased clusters in CF are related to CD8-positive, alpha-beta T cells, natural killer cells, and non-classical monocytes, suggesting a potential role for these cell types in the pathogenesis of CF. This is consistent with the known biology of these cell types, where they play essential roles in immune responses.

The differentially expressed genes in these clusters are related to cytokine production, VEGF signaling, and BCR signaling, suggesting a potential role for these pathways in the pathogenesis of CF. This is consistent with the known biology of these pathways, where they are essential for immune responses.

The hypoxia TXNIP NLRP3 inflammasome activation in CF lung are characterized by the upregulation of TXNIP-related genes such as TXNIP and NLRP3. This is consistent with the known biology of hypoxia TXNIP NLRP3 inflammasome activation, where this pathway is essential for inflammasome activation.

The VEGF signaling endothelial remodeling and angiogenesis in CF are characterized by the upregulation of VEGF-related genes such as VEGFA and TXNIP. This is consistent with the known biology of VEGF signaling, where this pathway is essential for angiogenesis.

The Oxidative Stress pathway is active in the respiratory tract suprabasal cell, suggesting a potential role for this pathway in the pathogenesis of CF. This is consistent with the known biology of Oxidative Stress, where this pathway is essential for oxidative stress responses.

The Fibrosis pathway is active in the fibroblast

Future Perspectives

Future perspectives for cystic fibrosis research based on ELISA analysis include:

- **Investigating the role of Type I IFN signaling in CF airway epithelium**: Further experiments should focus on elucidating the mechanisms by which Type I IFN signaling contributes to CF pathogenesis. This could involve analyzing the expression of IFN-stimulated genes and their downstream effects on cellular processes. Additionally, exploring the impact of Type I IFN signaling on epithelial defense mechanisms and its potential as a therapeutic target.
- **Examining the relationship between CD8 T cell activation and macrophage function in CF**: The interactions between CD8 T cells and macrophages in CF airway epithelium warrant further investigation. This could involve analyzing the expression of cytokines and chemokines involved in these interactions and their impact on disease progression. Additionally, exploring the role of immune checkpoints, such as HLA-E NKG2A, in regulating CD8 T cell activity.
- **Understanding the impact of hypoxia on TXNIP NLRP3 inflammasome activation in CF lung**: The role of hypoxia in activating the TXNIP NLRP3 inflammasome in CF lung tissue is an area of interest. Further experiments should focus on elucidating the mechanisms by which hypoxia regulates TXNIP expression and its downstream effects on inflammation and disease progression. Additionally, exploring the impact of TXNIP NLRP3 inflammasome activation on epithelial defense mechanisms and its potential as a therapeutic target.
- **Investigating the role of VEGF signaling in endothelial remodeling and angiogenesis in CF**: The involvement of VEGF signaling in endothelial remodeling and angiogenesis in CF lung tissue is an area of interest. Further experiments should focus on elucidating the mechanisms by which VEGF signaling regulates endothelial cell function and its impact on disease progression. Additionally, exploring the potential of VEGF signaling as a therapeutic target for CF treatment.

- **Exploring the relationship between chromatin remodeling and DNA damage repair in CF epithelium**: The impact of chromatin remodeling and DNA damage repair on CF epithelial cell function is an area of interest. Further experiments should focus on elucidating the mechanisms by which chromatin remodeling and DNA damage repair contribute to CF pathogenesis. Additionally, exploring the potential of targeting chromatin remodeling and DNA damage repair as a therapeutic strategy for CF treatment.

Methods

Dataset

The cystic fibrosis airway single-cell RNA-seq dataset was analyzed using ELISA.

Embedding Generation

Cluster-level embeddings were generated using a dual approach: (1) Semantic embeddings via BioBERT (pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb) encoding natural language cluster summaries including marker genes, GO terms, and Reactome pathways; (2) Expression embeddings via scGPT whole-human pre-trained model or PCA-based centroids as fallback.

Differential Expression

Wilcoxon rank-sum tests were performed on all genes (no HVG filtering) to identify cluster markers. Per-gene statistics (logFC, fraction expressing in/out of cluster) were stored for the top 2000 genes per cluster.

Comparative Analysis

Condition-specific gene expression differences were assessed by weighting cluster-level gene statistics by the proportion of cells from each condition within each cluster.

Cell-Cell Interaction Inference

Ligand-receptor interactions were predicted using a curated database derived from CellPhoneDB, CellChat, and KEGG. Interactions were scored by the product of ligand expression fraction in the source cluster and receptor expression fraction in the target cluster.

Cell Type Proportion Analysis

Cell type proportions were computed from cluster cell counts. Condition-specific proportions were estimated using metadata distribution weights.

Pathway Activity Scoring

Pathway activity was scored per cluster as the mean expression fraction (pct_in) of pathway member genes. Gene sets were derived from KEGG and Reactome databases.

LLM Interpretation

Retrieval results were interpreted using Llama-3.1-8B-Instant via Groq API, with strict grounding instructions to prevent hallucination.

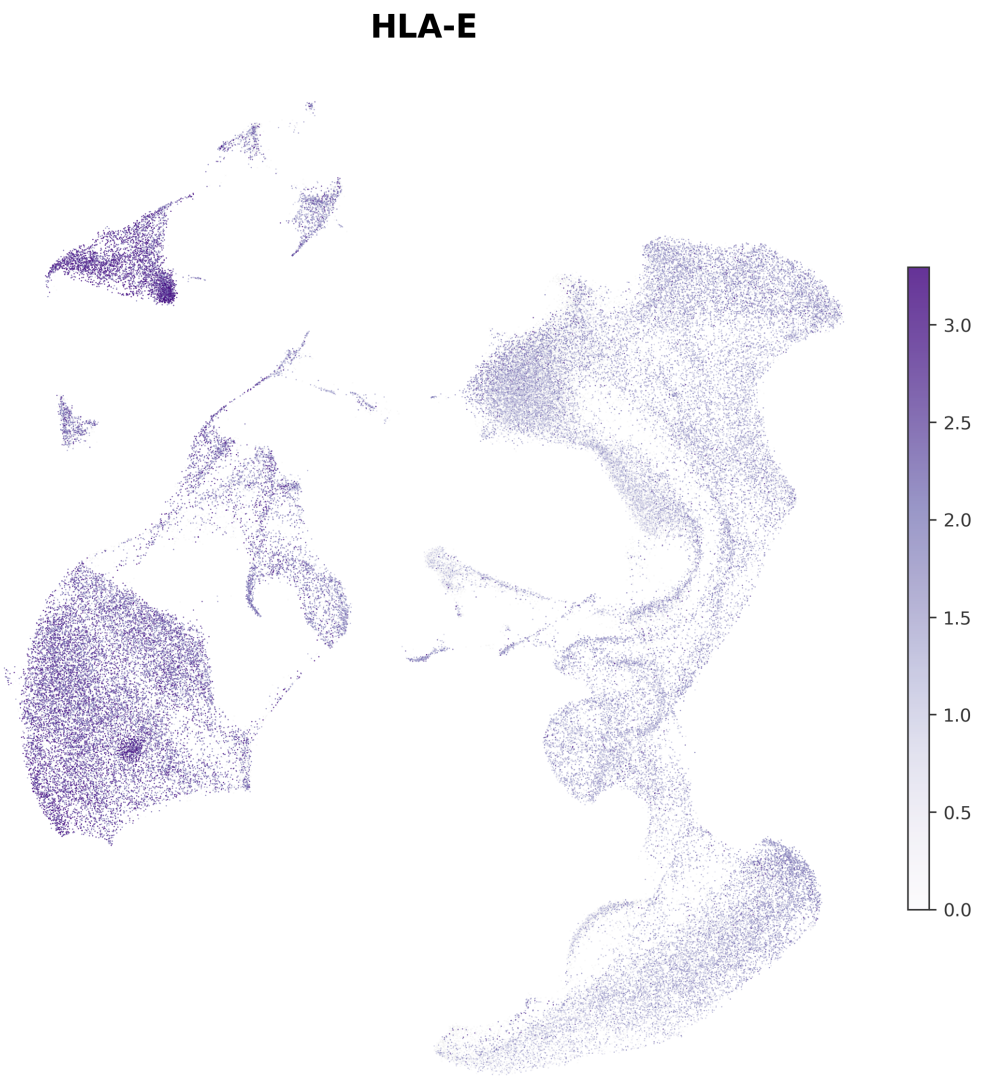

Figure 4. Expression of *HLA-E* projected onto the cell-level UMAP of the cystic fibrosis airway dataset (D1). Color intensity (purple gradient) indicates normalized expression level, with non-expressing cells shown in grey. *HLA-E* is most highly expressed in immune cell clusters, particularly $CD8^+$ T cells and NK cells, consistent with its role as a ligand for the NKG2A inhibitory receptor. Moderate expression is observed across epithelial populations including basal cells, supporting the HLA-E/NKG2A immune checkpoint axis identified by Berg *et al.*