Single-Cell Atlas of CD4+ T-Helper 2 Cells Reveals Distinct Cytokine Functions Across Diseases

Son Vo[1+], Maria Suprun[2+], Han Do[1], My Ngo[1], Rajesh Rao[2], Saurabh Vishnu Laddha[2], Marta E Polak[2], Calixte Monast[2*], Son Pham[1*].

1. BioTuring, 4445 Eastgate Mall, San Diego, California 92121, USA.
2. Johnson & Johnson, 1400 McKean Road, Spring House, PA, 19477, USA

*Corresponding author:
Son Pham, PhD, Calixte Monast, PhD.
Email: sonpham@bioturing.com, cmonast1@its.jnj.com

(+) Equal contributions.

Abstract

Despite the prevalence of autoimmune diseases, treatment options for specific conditions remain limited. These diseases are driven by three types of inflammatory pathways, each orchestrated by different T-helper cell subsets; however, a comprehensive atlas capturing their diverse functions across conditions and tissues is lacking. One of these subsets, $CD4^+$ T-helper type 2 (Th2) cells, is a relatively infrequent immune cell population that plays a central role in orchestrating type 2 inflammation. In this study, we developed the first comprehensive $CD4^+$Th2 single-cell atlas by analyzing the BioTuring database, which has data from over 300 million cells. This atlas integrates data from 52 studies, comprising a total of 39,243 cells derived from 9 major tissues, representing over 30 distinct diseases. Leveraging this atlas, we confirmed a rarely recognized $CD4^+$Th2 cell population co-expressing *IL22* and *IL13* observed exclusively in atopic dermatitis and not in other allergic conditions. By integrating single-cell data with other "omics" technologies, our analysis suggests distinct downstream effects of *IL22* and *IL13*, highlighting the potential for combination therapy to enhance treatment outcomes for AD patients. Additionally, by utilizing data derived from the wide range of diseases, tissues, and time points included in this atlas, we re-confirmed the unique connection between *IL9* expression and conditions related to allergen exposure. Pseudotime analysis detailed $CD4^+$Th2 cytokine dynamics upon allergen stimulation, showing a shift from early *IL2* and *IL4* expression to later *IL9* and *IL5*, with *IL13* consistently expressed, and further revealed that *IL9* production was transient. Collectively, these findings reveal distinct, context-dependent $CD4^+$Th2 cytokine functions, opening the potential for identifying novel therapeutic strategies for allergic and inflammatory diseases.

1. Introduction

$CD4^+$ T-helper type 2 (Th2) cells play a central role in type 2 immune responses, primarily targeting extracellular parasites. Current understanding of $CD4^+$Th2 biology largely revolves around well-characterized cytokines such as IL4, IL5, and IL13, which are crucial for recruiting eosinophils and basophils to sites of infection[1]. However, overactivation of $CD4^+$ Th2 cells has been implicated in chronic allergic inflammation, contributing to prevalent diseases like atopic dermatitis and asthma[2,3]. While existing therapies targeting these cytokines, such as dupilumab, have shown efficacy, not all patients achieve remission, suggesting involvement of additional, yet uncharacterized, factors/pathways in Th2-related diseases [4].

Despite their importance, the heterogeneity and functional diversity of $CD4^+$Th2 cells across tissues and disease conditions remain poorly understood. This challenge arises partly from the limited availability of markers that can reliably differentiate Th2 cells from other T-helper cell subsets, relying mostly on cytokine secretion[5]. In addition, the low frequency of CD4+Th2 cells in peripheral tissues further complicates their detection and analysis. Moreover, the plasticity of different T-helper subsets, which allows for transition between different classes, such as T follicular helper to CD4+Th2[6] or regulatory T cells to Th17[7], poses additional challenges for identifying classes when a T cell expresses more than one type of marker.

To address this gap, we constructed the first integrated CD4+Th2 single-cell atlas, encompassing diverse tissues and a wide range of conditions to systematically study CD4+Th2 heterogeneity. Through detailed analysis, we uncovered distinct tissue- and condition-specific properties of CD4+Th2 cells, identifying *IL22* and *IL9* as disease-specific factors involved in atopic dermatitis and allergen-driven conditions, respectively.

2. Results

2.1. An integrated CD4+ Th2 atlas

We constructed a CD4+Th2 atlas utilizing the BioTuring single-cell database, which has data for over 300M cells (Fig. 1a). First, to identify consensus marker genes for the CD4+Th2 population, we reviewed each study that characterized CD4+Th2 cells, following annotations provided by the authors. We compiled a comprehensive list of the marker genes used in these studies to differentiate CD4+Th2 cells from other subtypes, and classified a gene as a consensus CD4+Th2 marker if it was consistently used for Th2 annotation in more than three studies. To ensure specificity, we excluded marker genes that were used to annotate other cell types in more than three studies. The final CD4+Th2 marker panel included *IL4, IL5, IL9, IL13, IL17RB, PTGDR2, HPGDS*, and *IL1RL1*. A CD4+ T cell was classified as a Th2 subset cell if it expressed at least one of these markers and highly expressed GATA3. Using this panel, combined with markers from other cell types as the negative panel, we applied a cell-type prediction algorithm to identify CD4+Th2 cells in pre-existing individual datasets, which were then integrated to construct the atlas (see Methods for details).

Applying these criteria to the BioTuring single-cell database, which is derived from over 3,500 studies and has data from more than 300M cells, we identified 52 datasets with at least 50 CD4+Th2 cells represented in each (Supplementary Table 1). Over 2.8M T cells were extracted from selected datasets, however, only 39,243 were classified as CD4+Th2. We then visualized expression of our panel of CD4+Th2 marker genes on the integrated T-cell atlas, comparing the BioTuring-Th2-derived set with the authors' annotations (Fig. 1b). All the T cells expressed *PTPRC* and *CD3D*, which are established T-cell markers. While BioTuring-Th2 set showed expression of CD4+Th2 marker genes (*IL4, IL5, IL9, IL13, IL17RB, PTGDR2, HPGDS*, and *IL1RL1*) in the identified Th2 cell subset, there was minimal expression of these genes in other

T-cell subsets, indicating the specificity of our annotated atlas and confirming the rarity of the Th2 population (Fig 1b). The accuracy of our CD4+Th2 classification on different sequencing technologies is shown in Fig. S1a. The final CD4+Th2 atlas contains a set of 39,243 cells, encompassing 9 different major tissues and 9 major conditions, as well as healthy controls (Fig. 1c, d).

As shown in Fig. 1c, e, the atlas includes CD4+Th2 cells from a diverse range of diseases (e.g., allergy, asthma, cancer, infectious diseases, and skin inflammation) and healthy controls, as well as tissues (e.g., blood, skin, respiratory system, digestive system, kidney, cerebrospinal fluid, liver, and bone marrow). These broad categories encompass immunologically heterogeneous conditions with varying degrees of Th2 involvement. For example, among skin inflammatory diseases, Th2 cytokines (e.g., IL13) are prominently expressed in atopic dermatitis and prurigo nodularis, but not in Th17-driven conditions such as psoriasis[8]. Similarly, expression of Th2 markers varies across asthma subtypes and tumor types. To ensure immunological relevance, our downstream analyses focus on representative Th2-enriched contexts within each disease category. For clarity and consistency, we retain the use of broader category terms (e.g., asthma, cancer, skin inflammation) throughout the manuscript, with the understanding that they refer specifically to Th2-associated subsets, unless otherwise noted.

Some conditions had limited sampling sites; for example, all allergy samples were derived from blood, and all skin inflammatory samples came from skin tissue. Additionally, the atlas includes CD4+Th2 cells activated in vitro using established T-cell activation methods (PMID35618845). Beyond inflammatory and autoimmune diseases, the representation of CD4+Th2 cells derived from cancer was limited. Among the 11,330,316 T cells profiled from cancer samples in the BioTuring's single-cell database[9], only 2,761 (~0.024%) were classified as CD4+Th2 cells, indicating a limited role of this T-helper subset in cancer. Of the 39,243 CD4+Th2 cells

identified, only 682 (1.71%) originated from healthy samples, primarily from the skin and respiratory system, highlighting their rarity in non-disease conditions.

Following batch correction for study and assay technology, CD4+Th2 cells within the atlas tended to cluster based on their tissue of origin (Fig. 1c). CD4+Th2 cells from solid tissues clustered separately, with skin-derived CD4+Th2 cells being dominant and forming distinct clusters, suggesting potential tissue-specific properties of CD4+Th2 subsets.

2.2. Th2 Heterogeneity across Different Conditions

To explore Th2 heterogeneity across different diseases, we first visualized the expression of CD4+Th2 marker genes in major disease classes (Figure 2A). In allergy (peanut allergy, non-asthmatic subjects allergic to house dust mites) and asthma (allergic asthmatic patients, asthmatic subjects allergic to house dust mites), CD4+Th2 cells showed elevated expression of *IL4, IL5, IL9,* and *IL13*, highlighting their role in immune cell recruitment and inflammation[10–12]. In contrast, CD4+Th2 cells in Th2-associated inflammatory skin diseases predominantly expressed IL13, with minimal expression of the other three cytokines. We then visualized expression of the CD4+Th2 panel of markers across multiple studies, separated by tissue type. We observed that while *IL4, IL5,* and *IL13* showed expression in both blood and solid tissues, *IL9* was predominantly expressed in blood, with only minimal presence in solid tissues (Fig. S1b). In cancer, the expression levels of *IL4, IL5,* and *IL13* were low, while IL9 was undetectable. Upon investigation into more detailed conditions, we found that eosinophilic esophagitis (EoE), food allergy (FA), asthma/allergic respiratory diseases (A/ARD), and AD were the most extensively represented conditions in the atlas, based on cell numbers, samples, and datasets (Fig. 2b). As these were common inflammatory atopic diseases, these findings emphasize the central role of CD4+Th2 cells in inflammatory and allergic responses.

We then applied unsupervised clustering to identify sub-populations within the CD4+Th2 atlas (Fig. 2c), followed by differential gene expression analysis to extract top genes associated with each cluster. To focus on the immune regulatory roles of CD4+Th2, we narrowed the gene list (adjusted p-value < 0.05) to cytokines and chemokines. This approach revealed three major sub-populations (Fig. 2e). The first group (clusters 2, 3, 4, and 5) showed strong expression of *IL2RA, IL4, IL5, IL9*, and *IL13*. In contrast, cluster 6 represented a distinct group marked by elevated *IL22, IL26*, and *CCR10* expression, with minimal expression of cytokines from group 1. The third group (clusters 8 and 9) was characterized by a dominant chemokine profile, including *CXCR4 and CXCR6*. In terms of disease distribution, the first group was enriched in Kimura disease, EoE, FA, and A/ARD, whereas the second group was predominantly associated with AD.

To further understand the diversity in CD4+Th2 cell-driven inflammatory allergic mechanisms, we performed differential expression analysis comparing AD (n=9 datasets), A/ARD (n=4 datasets), FA (n=2 datasets), and EE (n=3 datasets) against normal control datasets (n=6). Given that CD4+Th2 cells mediate their functions through cytokines and their receptors, we focused on identifying cytokine- and receptor-related genes within the "KEGG CYTOKINE-RECEPTOR INTERACTION" pathway that are differentially expressed in these four conditions (Fig. 2f). Consistent with findings across major disease classes, *IL13* was elevated in CD4+Th2 cells in all conditions compared to controls. In contrast , *IL4, IL5*, and *IL9*, although previously defined as potential markers for CD4+Th2 cells, were predominantly expressed in FA, EoE, and A/ARD. Interestingly, AD did not exhibit high levels of expression of well-known Th2-associated cytokines such as *IL4* or *IL5,* but uniquely expressed *IL22* and *IL26*.

Due to the limited set of different tissue samples across these four diseases available in the atlas – predominantly skin in AD and blood in A/ARD, FA, and EE (Fig. 3a), we considered the

possibility that tissue type, rather than disease state, could be the primary driver of cytokine expression, in some cases. To investigate this, we examined *IL4, IL5*, and *IL9* expression across all diseases sampled from blood (Fig. 3b) and *IL22* and *IL26* expression across all diseases sampled from the skin (Fig. 3c) in the Th2 atlas. Apart from A/ARD, FA, and EE, only Kimura disease exhibited elevated *IL4, IL5,* and *IL9* expression, suggesting that not all the CD4+ Th2 cells in blood express these cytokines. Similarly, in addition to AD, expression of *IL22* and *IL26* was observed only in mycosis fungoides and prurigo nodularis. While prurigo nodularis also showed strong expression of these cytokines, the sample contained fewer than 100 cells, indicating the signal may stem from background expression. Collectively, these observations suggest that the elevated expression of *IL4, IL5, IL9, IL22*, and *IL26* is driven by disease type rather than the tissue of origin.

2.3. Th2 cells co-express IL22 and IL13 in AD

IL22 was uniquely upregulated in AD compared to healthy controls (Fig. 2f and 3c), with prominent co-expression of IL13 (Fig. 4a–c). To quantify this, we analyzed the number of cells expressing *IL13, IL22*, or both across all conditions (Fig. 4d). A cell was classified as expressing a gene if at least one raw count was detected in its expression profile. AD displayed a high percentage (23.81%, n=424) of cells co-expressing *IL13* and *IL22*, confirming this as a unique feature of AD. This pattern was consistently replicated in four different datasets in the atlas (Fig. 4e).

We next investigated the role of *IL22* in AD by comparing its expression in other types of inflammatory skin diseases. We observed that *IL22* was co-expressed with *IL17A* (a Th17 marker) in psoriasis (Fig. 5a). Despite their distinct immune response types and T-helper cell origins, both AD and psoriasis exhibited elevated *IL22* expression. Since both diseases share histological and clinical features such as epidermal hyperplasia and skin inflammation, we

hypothesized that *IL22* contributes to features of skin inflammation symptoms in AD independently of *IL13*.

We first performed differential gene expression (DEG) analysis comparing AD and normal skin samples. To avoid technical variation between datasets, we conducted DEG analyses separately for each dataset. In three datasets (GSE153760, GSE222840, and GSE230575), the top upregulated genes in AD were *KRT6A, KRT16, S100A8,* and *S100A9* (adjusted p-value < 0.05) (Fig. 5b). Visualization of these genes revealed consistent upregulation across all cell types in AD compared to normal samples, with the most pronounced differences observed in keratinocytes (Fig. S2)*.* To investigate the functional relevance of upregulation of these genes, we performed pathway enrichment analysis. AD samples exhibited upregulation of skin disorder pathways, including "Formation of the Cornified Envelope" and "Keratinization", both of which include *KRT6A* and *KRT16* genes. In contrast, *S100A8* and *S100A9* genes are associated with pathways related to "Antimicrobial Peptides" (Fig. 5c). While antimicrobial peptides are pertinent to the innate immune system, their excessive upregulation could contribute to skin inflammation[13]. Bonafice et al. 2005 showed by RT-PCR that *S100A7, S100A8,* and *S100A9* mRNA levels increased in a dose-dependent manner in response to *IL22*; western blot analysis confirmed that IL22 stimulation led to elevated levels of S100A proteins[14]. We then set out to evaluate the relationship between *IL22* and *KRT6A, KRT16, S100A8*, and *S100A9* to understand the connection between *IL22* and AD endotypes.

IL22 inhibition by fezakinumab, an IL22-blocking biologic agent, reduced immune activation in the skin, as evidenced by decreased expression of *S100* genes, especially in the IL22-high expression population (GSE99802[15]). However, this study did not address the effects of IL22 inhibition on *KRT6A* and *KRT16*. We then reanalyzed this dataset by comparing the expression patterns of these genes across patients treated with fezakinumab versus placebo (Fig. 5d).

Consistent with findings on *S100* genes from the original publication, stratification by baseline IL22 levels was critical for detecting meaningful transcriptional changes in *KRT* genes following treatment. In AD patients with high baseline IL22 expression, fezakinumab effectively reduced levels of *KRT6A* and *KRT16* expression after 4 weeks of treatment. However, no significant effect was observed in AD patients with low IL22 baseline levels or those receiving a placebo.

The relationship between *IL13* and four upregulated downstream genes (*KRT6A, KRT16, S100A8*, and *S100A9)* was mentioned in Bangert et al. 2021[16]. Treatment with dupilumab led to a significant reduction in the coverage and expression levels of *S100A8* and *S100A9* in keratinocytes (Fig. 5e). However, there were only minor changes in *KRT6A* and *KRT16* expressions. These results indicate that dupilumab effectively mitigates the upregulation of antimicrobial peptide gene expression, but does not impact expression of hyperplasia-related keratin genes.

Leveraging the atlas helped identify the unique pattern of *IL13* and *IL22* co-expression in AD. However, our analyses proposed distinct downstream effects on AD pathology. Dupilumab, which blocks *IL4RA*, primarily influences expression of antimicrobial peptide genes with no significant effect on keratinization genes. In contrast, fezakinumab, an IL22 blocker, reduced expression of both keratinization and antimicrobial peptide genes. Based on these observations, we suggest a combination therapy targeting both cytokines could provide more comprehensive benefits in managing AD .

2.4. Temporal expression of IL9 in CD4+Th2 upon stimulation

In contrast to *IL22* being predominantly expressed in AD, *IL9* was expressed only in FA, EE, and A/ARD (Fig. 2f). To determine whether this was due to the difference in tissue type, we

explored *IL9* expression in blood and bronchial tube samples across eight conditions, spanning allergen-induced to non-allergen-induced diseases (Fig. 6a). Both blood and bronchial tube samples from A/ARD exhibited high IL9 expression. In fact, all allergic conditions subjected to either natural or artificial stimulation expressed *IL9*, while non-allergic blood samples showed no *IL9* expression (adjusted p-value <0.05). This suggests that production of *IL9* in CD4+Th2 cells might be a consequence of allergen exposure.

To investigate *IL9* expression dynamics in Th2 cells, we performed pseudotime analysis on CD4+Th2 cells from contact allergic conditions in two datasets individually: blood samples from A/ARD patients stimulated with house dust mites for 6 hours (GSE146170, n=1,176 cells) and blood samples from EE patients stimulated with anti-CD3/CD28 for 6 hours (GSE175930, n=13,100 cells). Both analyses showed consistent interleukin expression patterns. Specifically, we found a transition from high *IL2* and *IL4* expression to high *IL9* and *IL5*, while *IL13* remained consistently expressed throughout the trajectory (Fig 6b, c). Given that *IL2* and *IL4* are typically associated with the early stages of CD4+Th2 polarization and activation, it is likely that *IL9* and *IL5* are expressed at later stages after stimulation.

To validate this hypothesis, we extended the analysis to include non-polarized CD4+ T cells, assuming pseudotime would progress from a naive to an activated CD4+Th2 state. Of the two datasets analyzed earlier, only the EE dataset (GSE175930) contained such cells. Therefore, we also included a bronchial tube asthma dataset (GSE193816) stimulated with house dust mite for 24 hours, despite the limited CD4+Th2 cell count (n = 373) in this dataset. Using non-polarized CD4+ T cells as the origin, both analyses revealed a temporal transition in CD4+Th2 cells from being *IL9* and *IL5* negative to becoming *IL9* and *IL5* positive, with consistent expression of *IL13* (Fig. 7a–d). This observation was consistent with the finding in the original study that *IL9* expression was elevated specifically in asthmatic airways upon allergen stimulation[17]. Notably, we observed early production of *IL2* in the EE and blood

datasets (Fig. 7c, d). Additionally, the bronchial tube asthma dataset showed a unique trajectory, with Th2 cells reverting from being *IL9* positive back to *IL9* negative at the pseudotime endpoint (Fig. 7d). When focusing on allergic datasets undergoing stimulation at different time points, we found a similar pattern: *IL9* expression increased initially but decreased over time (Fig. S4).

To better understand the mechanisms driving *IL9* expression upon allergen exposure, a pathway analysis was conducted using highly expressed genes from *IL9*+ and *IL9*− Th2 cells, comprising 7,298 and 12,569 cells, respectively, across all A/ARD, EE, and FA datasets. The analysis identified five pathways uniquely and significantly enriched in *IL9*+ Th2 cells, several of which are linked to inflammatory responses, reinforcing *IL9*'s established role in modulating immune responses in allergic diseases such as asthma (Fig. 7e). Intriguingly, *IL9*+ CD4+Th2 cells were also enriched for the PPAR signaling pathway, a critical regulator of systemic metabolism and energy homeostasis[18].

3. Discussion

This study leveraged a large-scale single-cell database to construct the first comprehensive atlas of human CD4+Th2 cells across diverse tissues and diseases. Addressing the known challenges of identifying and characterizing this relatively rare T cell subset, our integrated approach successfully identified nearly 40,000 Th2 cells, defined by a consensus marker panel. The atlas confirmed the scarcity of CD4+Th2 cells in healthy tissues and cancer, reinforcing their primary association with inflammatory and allergic conditions, particularly AD, EE, allergy, and asthma. Initial analysis revealed that CD4+Th2 cells cluster significantly by their tissue of origin, suggesting that tissue-specific imprinting influences their function, although our subsequent analyses pinpointed specific disease states as major drivers of distinct functional cytokine profiles.

This CD4+Th2 atlas contributes to our understanding of the diversity and specificity of this rare CD4+ T-cell subset. Further, this atlas provides a systematic, high-resolution view that enables comparisons across datasets and conditions, revealing context-dependent cytokine expression patterns that might otherwise be overlooked. Despite its relatively small size, the atlas has capacity to capture distinct pathogenic mechanisms across different tissues and disease states. Unsupervised clustering over the CD4+Th2 atlas revealed multiple sub-populations, each marked by unique cytokine and chemokine signatures. The atlas introduced novel observations, such as the independent downstream functions of *IL13* and *IL22* in atopic dermatitis, as well as the transient expression of *IL9* as a consequence of stimulation in allergy and asthma. As such, the Th2 atlas provides a resource for refining our understanding of Th2-mediated immunity and identifying new therapeutic opportunities in Th2-associated diseases.

4. Limitation of Study:

Although the creation of this CD4+Th2 atlas represents a pioneering effort in characterizing the complexities of this immune subset, certain limitations must be acknowledged. Derived consensus marker panels based on the majority of authors' CD4+Th2 annotations may introduce biases. The coverage across all tissue-condition combinations is uneven, with some diseases primarily sampled from single tissues. Furthermore, the rarity of CD4+Th2 cells, particularly in cancer and healthy states, limits the depth of analysis in these contexts. Specifically, our CD4+Th2 marker panel seems skewed toward activated CD4+Th2 cells, as these markers are primarily associated with cytokine secretion, potentially overlooking markers for memory CD4+Th2 cells. Inclusion of datasets from additional studies may help address these limitations in the future versions of this atlas.

5. Methods

5.1. Marker Gene Panels Identification for CD4+Th2

To accurately identify CD4+Th2 cells, we constructed a positive marker gene pool through an extensive review of scientific literature and single-cell data. In addition, we utilized BioTuring Venice[25] to extract Th2-specific markers from BioTuring database annotations. To ensure both consensus and specificity, we scored each marker gene based on the authors' Th2 annotations in existing datasets:

- True positive score: A gene was scored as true positive if it effectively defined the CD4+Th2 cluster without significant expression in other CD4+ T-cell subtypes across all datasets.
- False positive score: A gene was scored as false positive if it effectively defined the CD4+Th2 cluster in one dataset but showed significant expression in other cell types in other datasets.

We iteratively assessed all datasets containing CD4+Th2 cells, retaining genes with true positive scores greater than three and false positive scores less than three.

5.2. Seed Detection using AUCell

An AUCell[26] score was computed for every cell in the BioTuring single-cell database to detect CD4+Th2 cells (referred to as "seeds"). This score represents the enrichment of a given gene set in an individual cell. This scoring system was used instead of assessing expression of the exact gene panel to address the drop-out issue, which is common in single-cell sequencing technologies.

A combination to define each cell type includes:

combination = [positive panel, negative panel].

The positive panel of CD4+Th2 cells is defined as:

positive panel = [CD4+ T cell markers] + [Th2 markers]

CD4+ T cell markers include three genes: PTPRC, CD3D, and CD4. To address the gene drop-out issue, we created permutations of these markers. "PTPRC CD4" set was excluded, since it is used to detect myeloid cells instead of CD4+ T cells. Resulting three panels included***:***

1. PTPRC, CD3D, CD4
2. PTPRC, CD3D
3. CD3D, CD4

Following the approach above, *Th2 markers were separated into 9 permutation sets.* Finally, combining *[CD4+ T cell markers] and [Th2 markers]***,** we created a total of 3 x 9 = 27 positive panels for CD4+Th2 cell identification (table below).

We included one negative panel of cell markers. This included marker genes derived from other major cell types, such as Epithelial cells, Endothelial cells, Myeloid leukocytes, B cells, gamma-delta T cells, and CD8+ T cells, naive T cells, Th1, Th17, T follicular helper, and regulatory T cells.

Positive panel	**Negative panel**
PTPRC, CD3D, CD4, IL4	CD34, EPCAM, PECAM1, CD14, CD79A, CD8A, CD8B, TRDC, TRGC1, TRGC2, CD27, FCER1G, GZMA, LEF1, IFNG, CCR6, IL17A, IL17F, CXCR5, FOXP3
CD3D, CD4, IL4	
PTPRC, CD3D, IL4	
PTPRC, CD3D, CD4, IL5	
CD3D, CD4, IL5	

PTPRC, CD3D, IL5	
PTPRC, CD3D, CD4, IL9	
CD3D, CD4, IL9	
PTPRC, CD3D, IL9	
PTPRC, CD3D, CD4, GATA3-AS1	
CD3D, CD4, GATA3-AS1	
PTPRC, CD3D, GATA3-AS1	
PTPRC, CD3D, CD4, IL17RB	
CD3D, CD4, IL17RB	
PTPRC, CD3D, IL17RB	
PTPRC, CD3D, CD4, PTGDR2	
CD3D, CD4, PTGDR2	
PTPRC, CD3D, PTGDR2	
PTPRC, CD3D, CD4, IL13	
CD3D, CD4, IL13	
PTPRC, CD3D, IL13	
PTPRC, CD3D, CD4, HPGDS	
CD3D, CD4, HPGDS	
PTPRC, CD3D, HPGDS	
PTPRC, CD3D, CD4, IL1RL1	
CD3D, CD4, IL1RL1	
PTPRC, CD3D, IL1RL1	

Min-max scaling was applied on all cell type markers for each cell before AUCell calculation to ensure uniform capture rates for each gene when determining representative cell types.

Seeds detection algorithm:

Input:

expression_matrix: "min-max scaled" expression matrix of shape n_cells x n_genes

all_celltypes_genesets: Canonical marker geneset of each cell type. Each geneset contains multiple combinations of positive and negative panels.

all_celltypes_labels = [$celltype_1$, $celltype_2$,... , $celltype_M$]

all_celltypes_genesets: [$celltype_geneset_1$, $celltype_geneset_2$,... , $celltype_geneset_M$]

where

$celltype_geneset_M$ = [

[$positive_panel_1$, $negative_panel_1$],

[$positive_panel_2$, $negative_panel_2$],

...,

[$positive_panel_N$, $negative_panel_N$]

]

Output:

seed_labels: A dictionary mapping each *cell* to the cell type label

Procedure:

Function1: Geneset_Score (*cell_expression , positive_panel, negative_panel*):

```
final_score ← 0
for combination in celltype_geneset do:
    positive_panel_AUCell_score = AUCell ( positive_panel of that cell_expression )
    negative_panel_AUCell_score = AUCell ( negative_panel of that cell_expression )
    if negative_panel_AUCell_score == 0  and positive_panel_AUCell_score >= 0.8 then :
        final_score ← max ( [ final_score, positive_panel_AUCell_score ] )
    end if
end for
return final_score
End Function
```

Function2: Celltypes_Genesets_Score (*cell_expression , all_celltypes_labels , all_celltypes_genesets*):

```
celltypes_labels = [ ]
celltypes_scores = [ ]
for celltype_label and celltype_geneset  in  zip ( all_celltypes_labels, all_celltypes_genesets ) do:
    celltypes_labels ← celltypes_labels + [ celltype_label ]
    positive_panel = celltype_geneset [1]
    negative_panel = celltype_geneset [2]
    celltype_geneset_score =  Geneset_Score ( cell_expression , positive_panel ,  negative_panel )
    celltypes_scores ← celltypes_scores + [ celltype_geneset_score ]
end for

top_score = celltypes_scores [ argmax (celltypes_scores) ]
if top_score == 0 then:
    return "Unassigned"
else:
    return celltypes_labels [ argmax (celltypes_scores) ]
end if
End Function
```

```
Function3: Seeds_Detection (expression_matrix, all_celltypes_labels, all_celltypes_genesets )
    seed_labels = { }
    for cell_barcode_i in all cells do:
        cell_expression = expression_matrix [ i , : ]
        seed_labels [ cell_barcode_i ] ← Celltypes_Genesets_Score (
            cell_expression , all_celltypes_labels , all_celltypes_genesets
        )
    end for
    return seed_labels
End Function
```

5.3. Seed Refinement using Iterative PCA, KNN, and Louvain Clustering

In this refinement process, we smoothed the identified seeds to their closed neighbor cells, which were determined by an integrated pipeline of iterative Principal Component Analysis (PCA)[28], k-nearest neighbors (KNN)[29], and Louvain clustering[30]. This step helped to mitigate the drop-out effects prevalent in single-cell datasets, while also discerning and removing false-positive seeds resulting from background noise expression.

We executed the pipeline independently within each sample across all studies, without employing batch correction methods. When performing PCA and Louvain clustering to annotate the cell types, clusters were correctly annotated only when all of its sub-clusters exhibited similar annotations. Any divergence from this pattern necessitated separation into smaller clusters to ensure accurate cell type annotations, the Louvain clustering algorithm was executed at high resolution to smooth the seeds effectively.

Given the focus on CD4+Th2 cells, all cell types apart from T cells resulting from seed detection were labeled as "Unassigned". Throughout the smoothing process, clusters consisting solely of unassigned cells were iteratively eliminated, and PCA and the Louvain algorithm were rerun to

improve separation between CD4+ T cell classes, thus improving Th2 annotation. A cluster was annotated as CD4+Th2 if its cell count ratio exceeded that of all other T cells. Finally, all the Th2 clusters from each of the different studies in the database were merged to create the final atlas.

PCA:

n_Principal_components = min(50, n_cells - 1).

k-NN: n_neighbors = 30.

Louvain clustering:

resolution = iterative from 1 -> 3, step=0.2; stop when n_clusters = 3 * n_celltypes_classes.

Assign CD4+ sub-cell type to cluster:

Input:
clusters_celltypes: A dictionary mapping cluster to its cell types' seed counts
clusters_celltypes = {
  $cluster_1$: $cluster_info_1$,
  $cluster_2$: $cluster_info_2$,
  ...,
  $cluster_M$: $cluster_info_M$,
}
where $cluster_info_M$: {
  labels : [$celltype_1$, $celltype_2$, ..., $celltype_N$] ,
  seed_counts : [n_Unassigned, $n_celltype_1$, $n_celltype_2$, ..., $n_celltype_N$]
}

Output:
cluster_celltype_result: A dictionary mapping each *cluster_i* to its assigned cell type label

Procedure:
cluster_celltype_result = {}
for $cluster_i$ **in** *clusters_celltypes* **do**:
  all_celltypes_labels = *clusters_celltypes* [$cluster_i$] [*labels*]
  all_celltypes_counts = *clusters_celltypes* [$cluster_i$] [*seed_counts*]
  total_cell = $\sum$ (*all_celltypes_counts*)
  unassigned_proportions = $\frac{n_Unassigned}{total_cell}$

  if *unassigned_proportions* > 0.7 **then:**
    cluster_celltype_result [$cluster_i$] ← "*Unassigned*"
    Proceed to the next iteration

end if

$proportions\ [j] = \frac{n_celltype\ [j]}{total_cell - n_Unassigned}$ for $j = 1, \ldots, N$
top_celltype = **argmax** (*proportions*)
2nd_top_celltype = **argmax** (*proportions [j]* for $j \neq$ *top_celltype*)
difference_between = proportions [top_celltype] - proportions [2nd_top_celltype]

if *top_celltype* > 0.5 **and** *difference_between* > 0.2 ***then:***
cluster_celltype_result [cluster$_i$ *]* ← *all_celltypes_labels [top_celltype]*
else:
cluster_celltype_result [cluster$_i$ *]* ← *"Unassigned"*
end if
end for

5.4. Preprocessing and Integrating for Atlas using scVI

Each selected dataset underwent individual preprocessing procedures strictly following the filtering parameters defined by the authors in the original papers (supplement table). All datasets included in the atlas were collected as raw integer counts, and the samples with fewer than 50 CD4+Th2 cells were excluded. The sequencing depth of all cells was then normalized by dividing the total counts over all genes of each cell by 10,000 and taking the natural logarithm of these values .

To ensure consistency of metadata terms across multiple datasets, the metadata of each dataset collected from each original publication, were carefully standardized based on EBI ontologies. For batch integration, we initially identified 1,000 highly variable genes (HVGs) using the seurat_v3 package implemented in Scanpy[31]. Subsequently, scVI[32] was applied to the raw counts of this set of 1,000 HVGs. The integration process involved treating “Study ID” variable as the batch variable. The scVI method was run using specific parameter configurations: number of hidden dimensions: 128; number of layers: 1; number of latent dimensions: 30; gene likelihood: zero-inflated negative binomial (zinb; batch size: 128 and kl_weight: 0.1. The model

was trained with 100 epochs to avoid over-batch correction, and all other parameters were set to default values. Finally, we computed a k-nearest neighbors graph with n_neighbor = 90 and applied UMAP[33] using the scVI latent space for visualization.

5.5. Differential gene expression and Pathway analysis

Unless otherwise stated, Differential Gene Expression (DGE) analysis was performed using the BioTuring BBrowserX platform with the Venice statistical method. The analysis criteria included adjusted p-values < 0.05, |log2FC| > 0.5, Bonferroni correction for multiple comparisons, and a default min.pct threshold of 0.1 for gene coverage. The resulting gene scores were used as input for Decoupler[34], while the differential gene list was put into GSEApy[35] for pathway enrichment analysis. Enrichment tests incorporated pathways from the MsigDB[36] (MSigDB_Hallmark_2020), KEGG[37] (KEGG_2021_Human), GO[38,39] (GO_Biological_Process_2023), and Reactome databases[40] (Reactome_Pathways_2024). Pathways were considered significant if they had a threshold of adjusted p-values of < 0.05.

5.6. Trajectory analysis

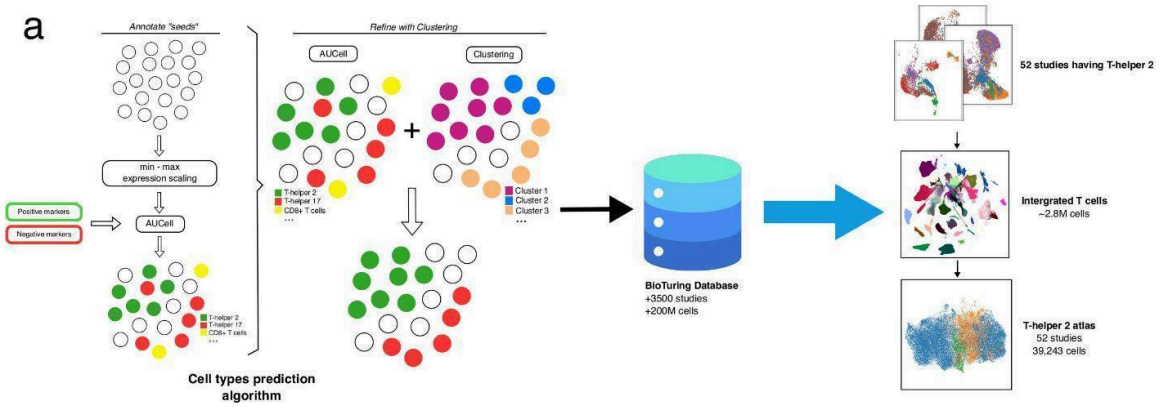
Pseudotime trajectory analysis was performed using Monocle3 to investigate the temporal dynamics of interleukin expression in CD4+Th2 cells during allergen stimulation. To minimize batch effects from variations in condition, sequencing technology, and tissue, the analysis was performed on each dataset independently. Non-polarized CD4+ T cells were selected as the starting point to establish a naive-to-activated trajectory.

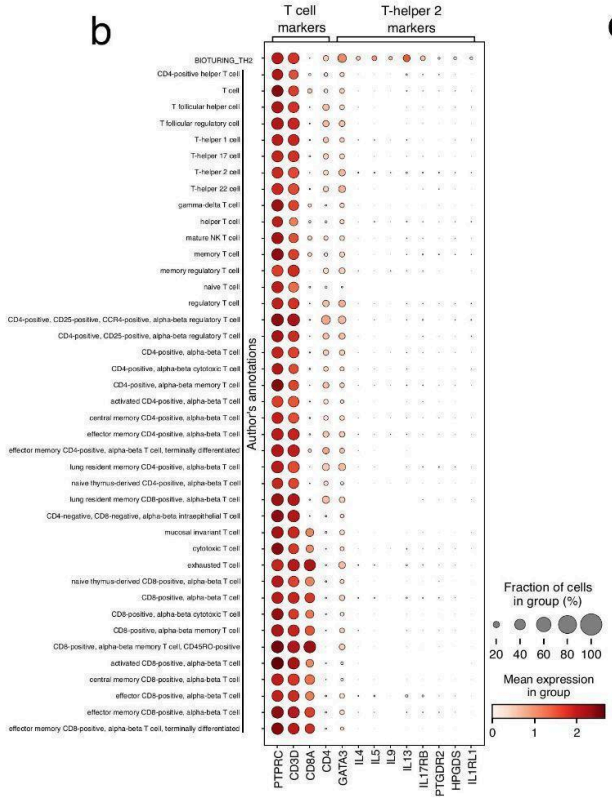
Acknowledgements

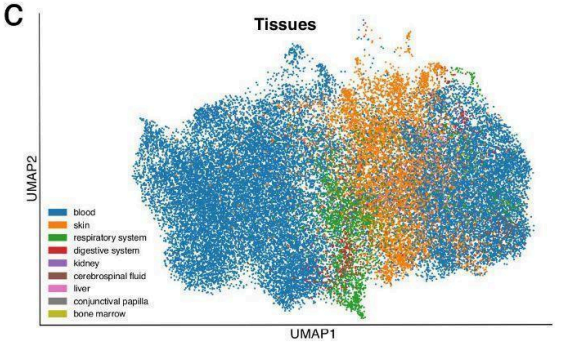
Author contributions:

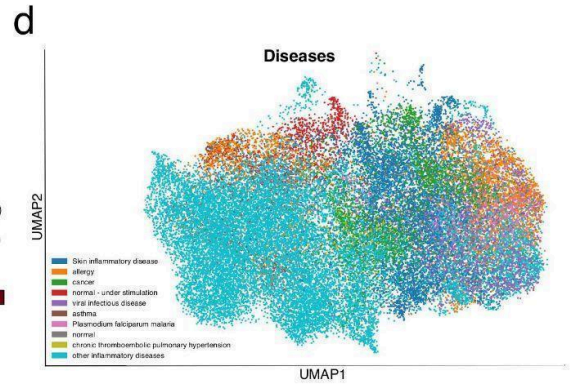
SV created the pipeline, analyzed the data, prepared the figures, and wrote the manuscript. MN analyzed the data and wrote the manuscript. HD analyzed the data. MS, RR, SL, MP wrote the manuscript, designed the analytics. SP, CM conceived of the study, contributed to the creation of the algorithm, writing of the manuscript, and supervised the study.

Figures


a
Annotate "seeds"
min - max expression scaling
Positive markers
Negative markers
AUCell
T-helper 2
T-helper 17
CD8+ T cells
Refine with Clustering
AUCell
Clustering
Cluster 1
Cluster 2
Cluster 3
Cell types prediction algorithm
BioTuring Database
+3500 studies
+200M cells
52 studies having T-helper 2
Intergrated T cells
~2.8M cells
T-helper 2 atlas
52 studies
39,243 cells


b
T cell markers
T-helper 2 markers
Author's annotations
PTPRC
CD3D
CD8A
CD4
GATA3
IL4
IL5
IL9
IL13
IL17RB
PTGDR2
HPGDS
IL1RL1
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0 1 2


c
Tissues
UMAP2
UMAP1
blood
skin
respiratory system
digestive system
kidney
cerebrospinal fluid
liver
conjunctival papilla
bone marrow


d
Diseases
UMAP2
UMAP1
Skin inflammatory disease
allergy
cancer
normal - under stimulation
viral infectious disease
asthma
Plasmodium falciparum malaria
normal
chronic thromboembolic pulmonary hypertension
other inflammatory diseases

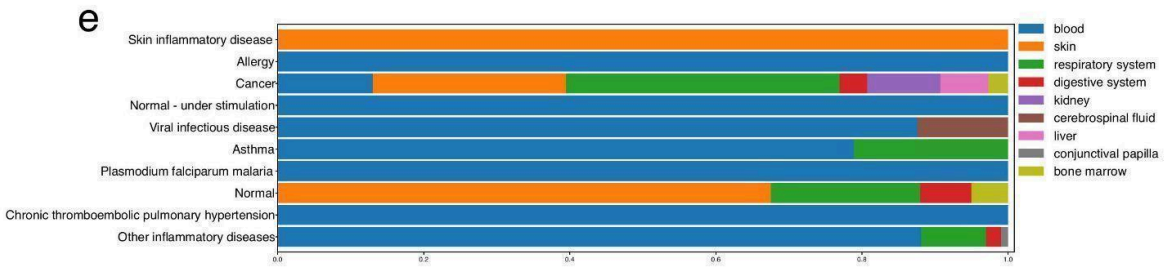

e
Skin inflammatory disease
Allergy
Cancer
Normal - under stimulation
Viral infectious disease
Asthma
Plasmodium falciparum malaria
Normal
Chronic thromboembolic pulmonary hypertension
Other inflammatory diseases
0.0
0.2
0.4
0.6
0.8
1.0
blood
skin
respiratory system
digestive system
kidney
cerebrospinal fluid
liver
conjunctival papilla
bone marrow

Figure 1. CD4+Th2 atlas overview

a. Summary of the cell type prediction algorithm and diagram of building the CD4+Th2 atlas from the BioTuring single-cell database.

b. Dotplot visualizing expression of T cell marker genes and consensus Th2 marker genes (in columns) between BioTuring's CD4+Th2 and the authors' annotation (in rows) in the CD4+Th2 cell atlas.

c. Distribution of major tissue types across CD4+Th2 cell atlas on UMAP.

d. Distribution of conditions across CD4+Th2 cell atlas

e. Proportions of tissues of origin in each condition

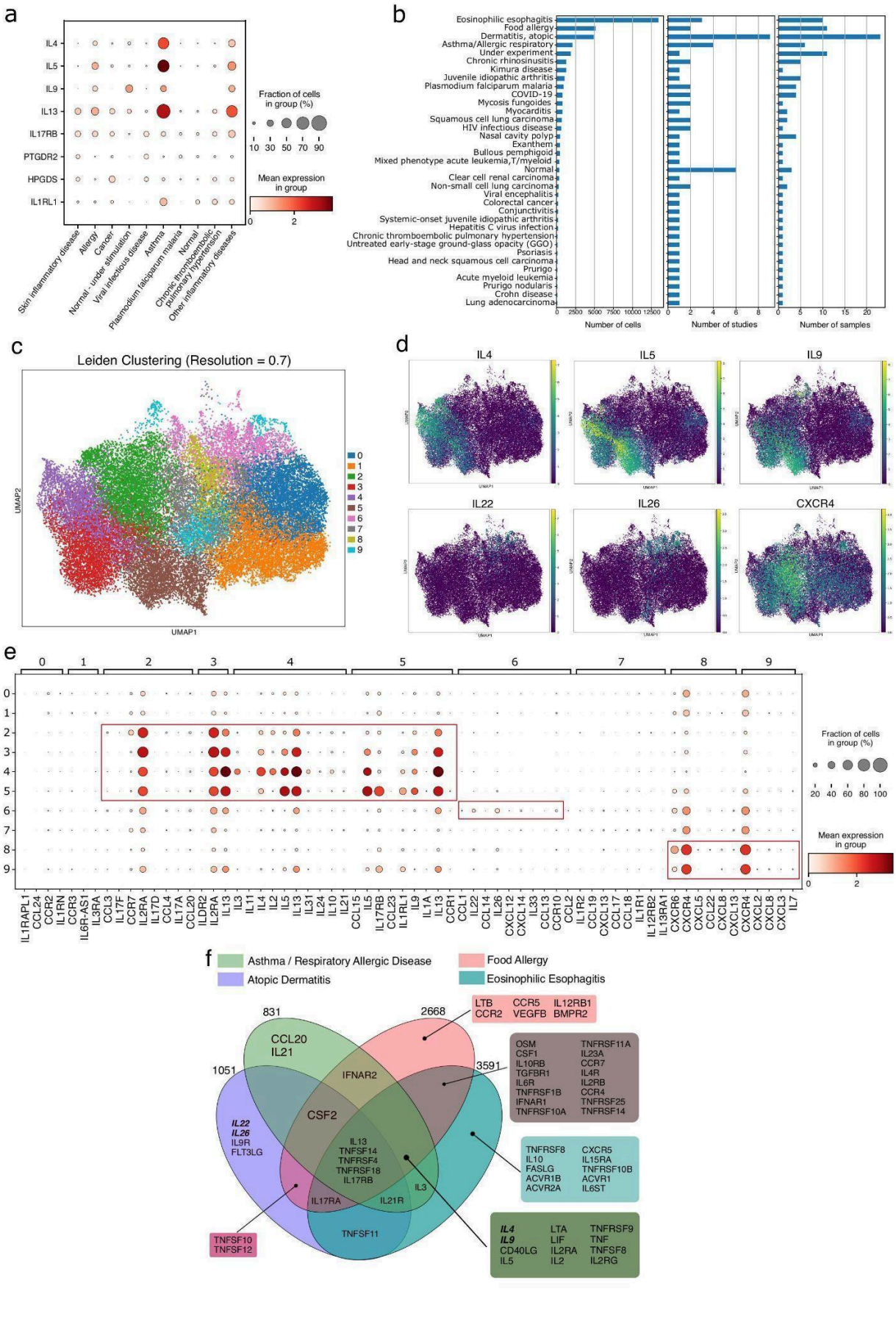

a
IL4
IL5
IL9
IL13
IL17RB
PTGDR2
HPGDS
IL1RL1
Skin inflammatory disease
Allergy
Cancer
Normal - under stimulation
Viral infectious disease
Asthma
Plasmodium falciparum malaria
Normal
Chronic thromboembolic pulmonary hypertension
Other inflammatory diseases
Fraction of cells in group (%)
10 30 50 70 90
Mean expression in group
0 2
b
Eosinophilic esophagitis
Food allergy
Dermatitis, atopic
Asthma/Allergic respiratory
Under experiment
Chronic rhinosinusitis
Kimura disease
Juvenile idiopathic arthritis
Plasmodium falciparum malaria
COVID-19
Mycosis fungoides
Myocarditis
Squamous cell lung carcinoma
HIV infectious disease
Nasal cavity polyp
Exanthem
Bullous pemphigoid
Mixed phenotype acute leukemia,T/myeloid
Normal
Clear cell renal carcinoma
Non-small cell lung carcinoma
Viral encephalitis
Colorectal cancer
Conjunctivitis
Systemic-onset juvenile idiopathic arthritis
Hepatitis C virus infection
Chronic thromboembolic pulmonary hypertension
Untreated early-stage ground-glass opacity (GGO)
Psoriasis
Head and neck squamous cell carcinoma
Prurigo
Acute myeloid leukemia
Prurigo nodularis
Crohn disease
Lung adenocarcinoma
0 2500 5000 7500 10000 12500
Number of cells
0 2 4 6 8
Number of studies
0 5 10 15 20
Number of samples
c
Leiden Clustering (Resolution = 0.7)
UMAP2
UMAP1
0
1
2
3
4
5
6
7
8
9
d
IL4
IL5
IL9
IL22
IL26
CXCR4
e
0 1 2 3 4 5 6 7 8 9
IL1RAPL1 CCL24 CCR2 IL1RN CCR3 IL6R-AS1 IL3RA CCL3 IL17F CCR7 IL2RA IL17D CCL4 IL17A CCL20 ILDR2 IL2RA IL13 IL3 IL11 IL4 IL2 IL5 IL13 IL31 IL24 IL10 IL21 CCL15 IL5 IL17RB CCL23 IL1RL1 IL9 IL1A IL13 CCR1 CCL1 IL22 CCL14 IL26 CXCL12 CXCL14 IL33 CCL13 CCR10 CCL2 IL1R2 CCL19 CXCL13 CCL17 CCL18 IL1R1 IL12RB2 IL13RA1 CXCR6 CXCR4 CXCL5 CCL22 CCL8 CXCL13 CXCR4 CXCL2 CXCL8 CXCL3 IL7
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0 2
f
Asthma / Respiratory Allergic Disease
Food Allergy
Atopic Dermatitis
Eosinophilic Esophagitis
831
2668
1051
3591
CCL20
IL21
IFNAR2
CSF2
IL22
IL26
IL9R
FLT3LG
IL13
TNFSF14
TNFRSF4
TNFRSF18
IL17RB
IL3
IL17RA
IL21R
TNFSF11
TNFSF10
TNFSF12
LTB CCR5 IL12RB1
CCR2 VEGFB BMPR2
OSM TNFRSF11A
CSF1 IL23A
IL10RB CCR7
TGFBR1 IL4R
IL6R IL2RB
TNFRSF1B CCR4
IFNAR1 TNFRSF25
TNFRSF10A TNFRSF14
TNFRSF8 CXCR5
IL10 IL15RA
FASLG TNFRSF10B
ACVR1B ACVR1
ACVR2A IL6ST
IL4 LTA TNFRSF9
IL9 LIF TNF
CD40LG IL2RA TNFSF8
IL5 IL2 IL2RG

Figure 2. CD4+Th2 gene expression profiles across different conditions

a. Expression of CD4+Th2 cell markers between different groups of conditions, with enrichment of interleukin genes in asthma, allergy, and other inflammatory diseases.

b. Number of cells, studies, and samples included for each condition in the CD4+Th2 cell atlas.

c. Unsupervised Leiden clustering with resolution=0.7 on scVI embedding.

d Expression of *IL4, IL5, IL9, IL22, IL26*, and *CXCR4* were visualized on UMAP.

e. Dotplot showing the expression and coverage of the top differentially expressed cytokines and chemokines of each cluster.

f. Venn diagram showing the overlap of differentially expressed genes in different Th2-associated diseases compared to healthy controls. Genes shown in the diagram are from the "KEGG CYTOKINE-RECEPTOR INTERACTION" pathway.

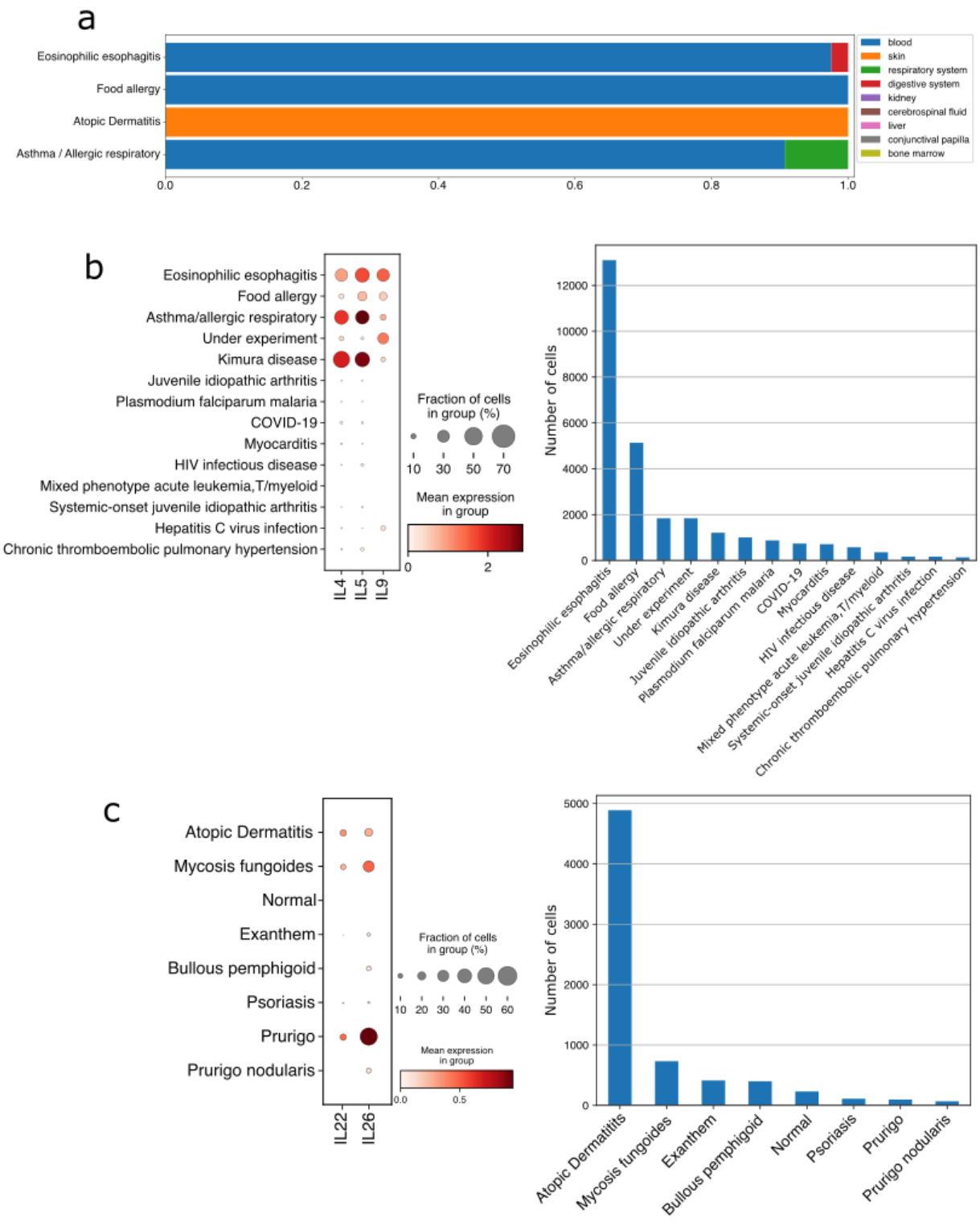

a
Eosinophilic esophagitis
Food allergy
Atopic Dermatitis
Asthma / Allergic respiratory
0.0
0.2
0.4
0.6
0.8
1.0
blood
skin
respiratory system
digestive system
kidney
cerebrospinal fluid
liver
conjunctival papilla
bone marrow
b
Eosinophilic esophagitis
Food allergy
Asthma/allergic respiratory
Under experiment
Kimura disease
Juvenile idiopathic arthritis
Plasmodium falciparum malaria
COVID-19
Myocarditis
HIV infectious disease
Mixed phenotype acute leukemia,T/myeloid
Systemic-onset juvenile idiopathic arthritis
Hepatitis C virus infection
Chronic thromboembolic pulmonary hypertension
IL4
IL5
IL9
Fraction of cells in group (%)
10
30
50
70
Mean expression in group
0
2
Number of cells
0
2000
4000
6000
8000
10000
12000
Eosinophilic esophagitis
Food allergy
Asthma/allergic respiratory
Under experiment
Kimura disease
Juvenile idiopathic arthritis
Plasmodium falciparum malaria
COVID-19
Myocarditis
HIV infectious disease
Mixed phenotype acute leukemia, T/myeloid
Systemic-onset juvenile idiopathic arthritis
Hepatitis C virus infection
Chronic thromboembolic pulmonary hypertension
c
Atopic Dermatitis
Mycosis fungoides
Normal
Exanthem
Bullous pemphigoid
Psoriasis
Prurigo
Prurigo nodularis
IL22
IL26
Fraction of cells in group (%)
10
20
30
40
50
60
Mean expression in group
0.0
0.5
Number of cells
0
1000
2000
3000
4000
5000
Atopic Dermatitis
Mycosis fungoides
Exanthem
Bullous pemphigoid
Normal
Psoriasis
Prurigo
Prurigo nodularis

Figure 3. Cytokine expression in CD4+ Th2 cell atlas is disease-driven rather than tissue-dependent.

a. Distribution of tissues in EE, FA, AD, and A/ARD.

b. Expression levels of *IL4*, *IL5,* and *IL9* in all diseases sampling from blood in the Th2 cell atlas. Elevated expression was only observed in A/ARD, FA, EE, and Kimura disease, suggesting these cytokines are not associated with blood-derived Th2 cells.

c. Expression of *IL22* and *IL26* in all diseases sampling from skin in the Th2 cell atlas. Beyond AD, notable expression was observed only in mycosis fungoides and prurigo nodularis, but with low cell counts, making it difficult to distinguish from background signals. This observation supports a disease-specific, rather than tissue-specific, pattern of *IL22* and *IL26* in CD4+Th2 cells.

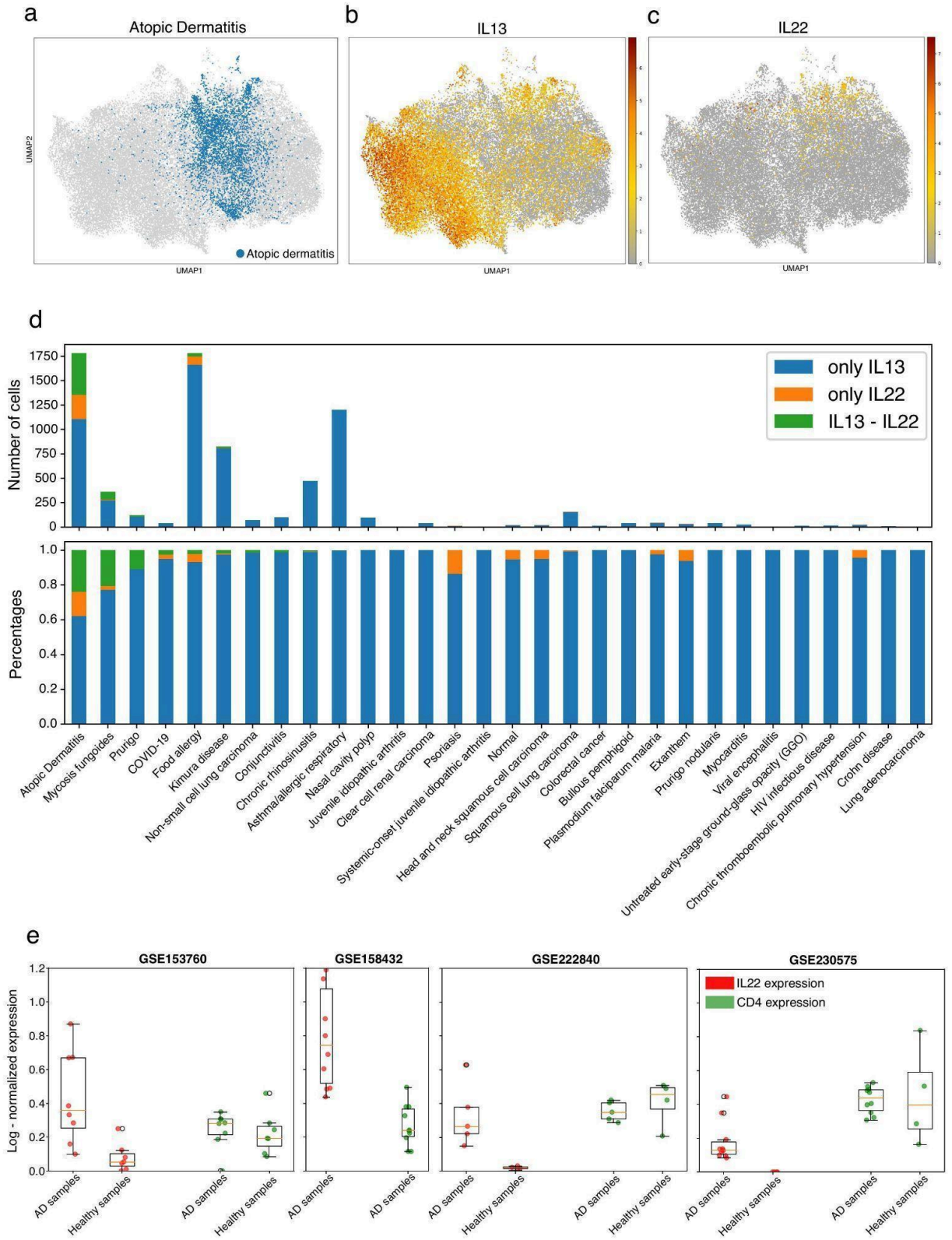

a
Atopic Dermatitis
UMAP2
UMAP1
Atopic dermatitis
b
IL13
UMAP1
c
IL22
UMAP1
d
Number of cells
Percentages
only IL13
only IL22
IL13 - IL22
Atopic Dermatitis
Mycosis fungoides
Prurigo
COVID-19
Food allergy
Kimura disease
Non-small cell lung carcinoma
Conjunctivitis
Chronic rhinosinusitis
Asthma/allergic respiratory
Nasal cavity polyp
Juvenile idiopathic arthritis
Clear cell renal carcinoma
Psoriasis
Systemic-onset juvenile idiopathic arthritis
Normal
Head and neck squamous cell carcinoma
Squamous cell lung carcinoma
Colorectal cancer
Bullous pemphigoid
Plasmodium falciparum malaria
Exanthem
Prurigo nodularis
Myocarditis
Viral encephalitis
Untreated early-stage ground-glass opacity (GGO)
HIV infectious disease
Chronic thromboembolic pulmonary hypertension
Crohn disease
Lung adenocarcinoma
e
GSE153760
GSE158432
GSE222840
GSE230575
Log - normalized expression
IL22 expression
CD4 expression
AD samples
Healthy samples

Figure 4. IL22 and IL13 Coexpression in Atopic Dermatitis.

a. Distribution of cells from atopic dermatitis disease, along with *IL13* **(b)** and *IL22* **(c)** expression embedded on the same UMAP. Co-expression of *IL13* and *IL22* was primarily observed in atopic dermatitis.

d. Number (top) and percentage (bottom) of cells expressing *IL13* and *IL22* across conditions. Atopic Dermatitis showed the largest number and ratio of co-expressed cells.

e. Mean expression (log-normalized values) of *IL22* (red) and *CD4* (green) in CD4+ T cells in four studies that included atopic dermatitis ; each dot represents an individual batch/sample. All studies show a consistently elevated level of *IL22* expression in AD compared to healthy controls, while *CD4* expression, as a control, remained similar between AD and healthy controls.

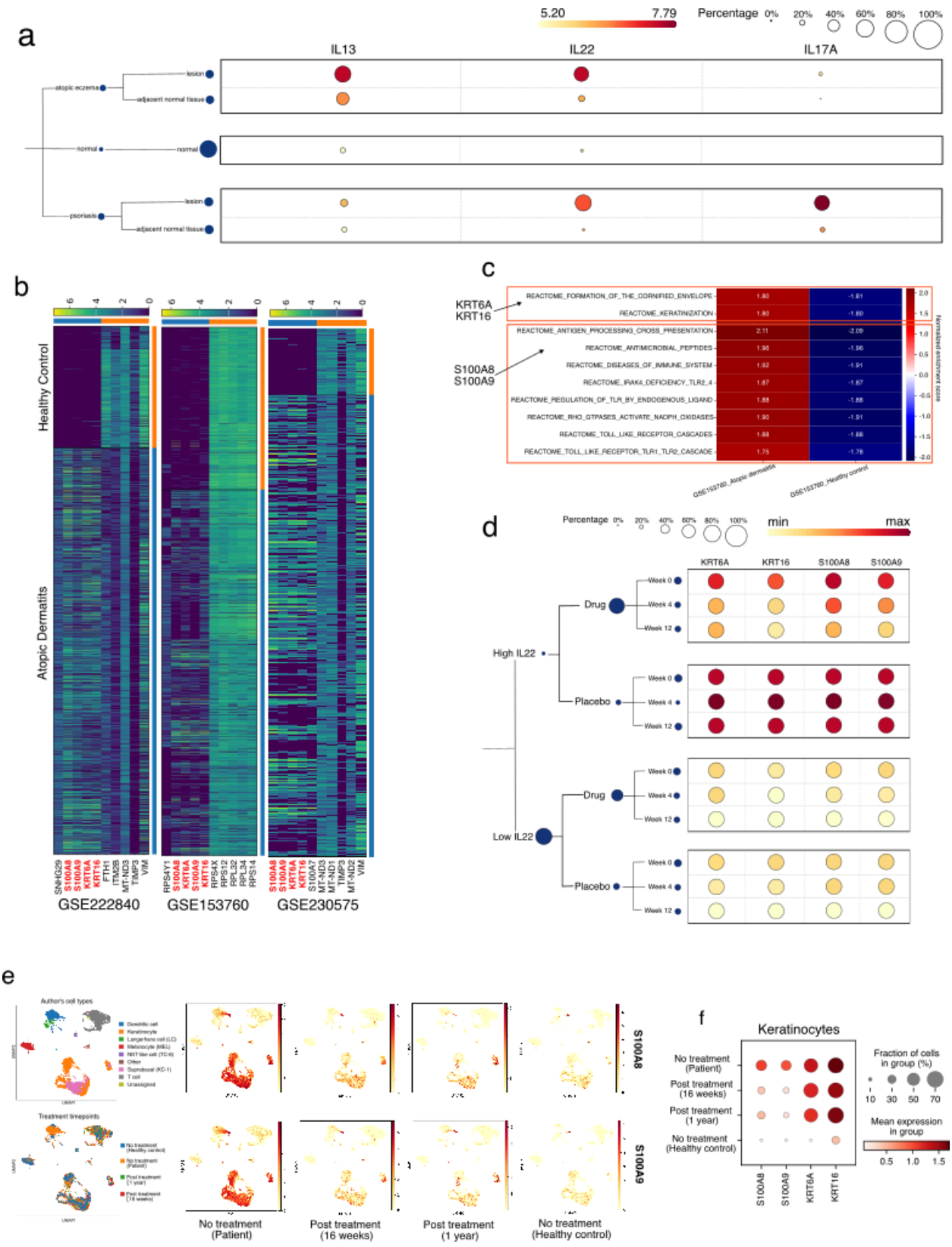

a
5.20
7.79
Percentage
0%
20%
40%
60%
80%
100%
IL13
IL22
IL17A
atopic eczema
lesion
adjacent normal tissue
normal
psoriasis
b
Healthy Control
Atopic Dermatitis
GSE222840
GSE153760
GSE230575
c
KRT6A
KRT16
S100A8
S100A9
REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE
REACTOME_KERATINIZATION
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION
REACTOME_ANTIMICROBIAL_PEPTIDES
REACTOME_DISEASES_OF_IMMUNE_SYSTEM
REACTOME_IRAK4_DEFICIENCY_TLR2_4
REACTOME_REGULATION_OF_TLR_BY_ENDOGENOUS_LIGAND
REACTOME_RHO_GTPASES_ACTIVATE_NADPH_OXIDASES
REACTOME_TOLL_LIKE_RECEPTOR_CASCADES
REACTOME_TOLL_LIKE_RECEPTOR_TLR1_TLR2_CASCADE
GSE153760_Atopic dermatitis
GSE153760_Healthy control
Normalized enrichment score
d
Percentage
min
max
KRT6A
KRT16
S100A8
S100A9
High IL22
Low IL22
Drug
Placebo
Week 0
Week 4
Week 12
e
Author's cell types
Treatment timepoints
No treatment (Patient)
Post treatment (16 weeks)
Post treatment (1 year)
No treatment (Healthy control)
S100A8
S100A9
f
Keratinocytes
Fraction of cells in group (%)
Mean expression in group

Figure 5: Downstream function of IL22 and IL13 in Atopic Dermatitis.
a. Expression of *IL13, IL22,* and *IL17A* in atopic dermatitis, psoriasis, and healthy controls. *IL22* was still expressed when lacking either *IL13* (in psoriasis) or *IL17A* (in AD), proving its independent roles.
b. Differential gene expression analyses between AD patients and healthy controls; all three studies showed the same top-upregulated genes in AD: *KRT6A, KRT16, S100A8*, and *S100A9*.
c. GSEA heatmap showing the upregulated keratinization pathway related to *KRT6A, KRT16, S100A8,* and *S100A9* in AD patients compared to healthy controls.
d. Fezakinumab reduces levels of *KRT* and *S100A* gene expression in atopic dermatitis patients with high IL22 baseline.
e, f. Dupilumab reduces S100A gene expression and has a minor effect on KRT genes in keratinocytes from atopic dermatitis patients.

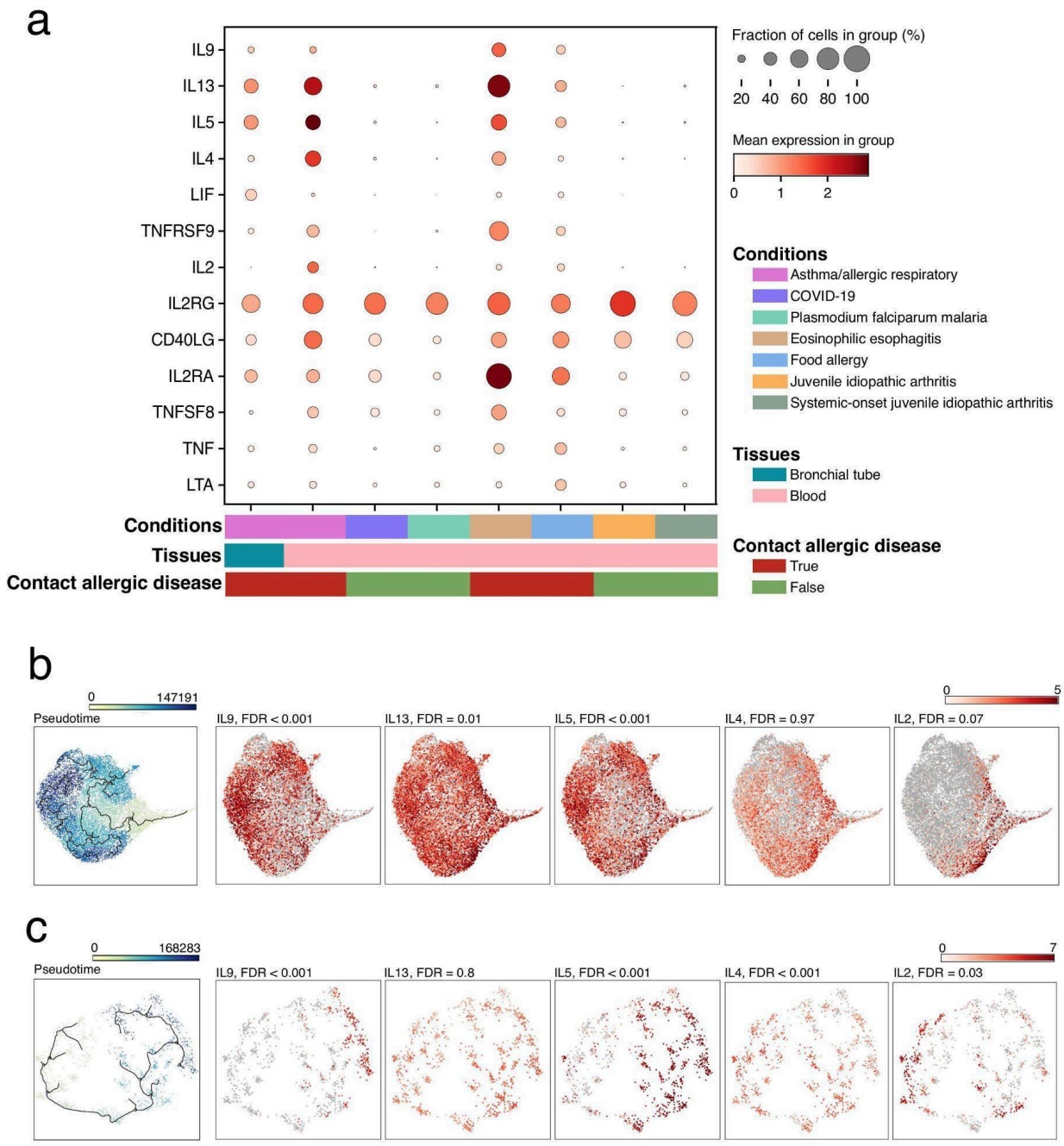


Figure 6. IL9+ CD4+Th2 Subpopulation Characterization.
a. Dotplot showing upregulation of genes across different conditions in bronchial tube and blood samples.
b, c. Pseudotime trajectory analysis and expression patterns of *IL9, IL13, IL5, IL4*, and *IL2* in **(b)** eosinophilic esophagitis (EE) (GSE175930, n = 13,100 Th2 cells) and **(c)** A/ARD (GSE146170, n = 1,176 Th2 cells). In both diseases, *IL13* maintained a consistent expression pattern throughout the trajectory, while *IL9* and *IL5* were enriched at one end and *IL4* and *IL2* at the opposite end.

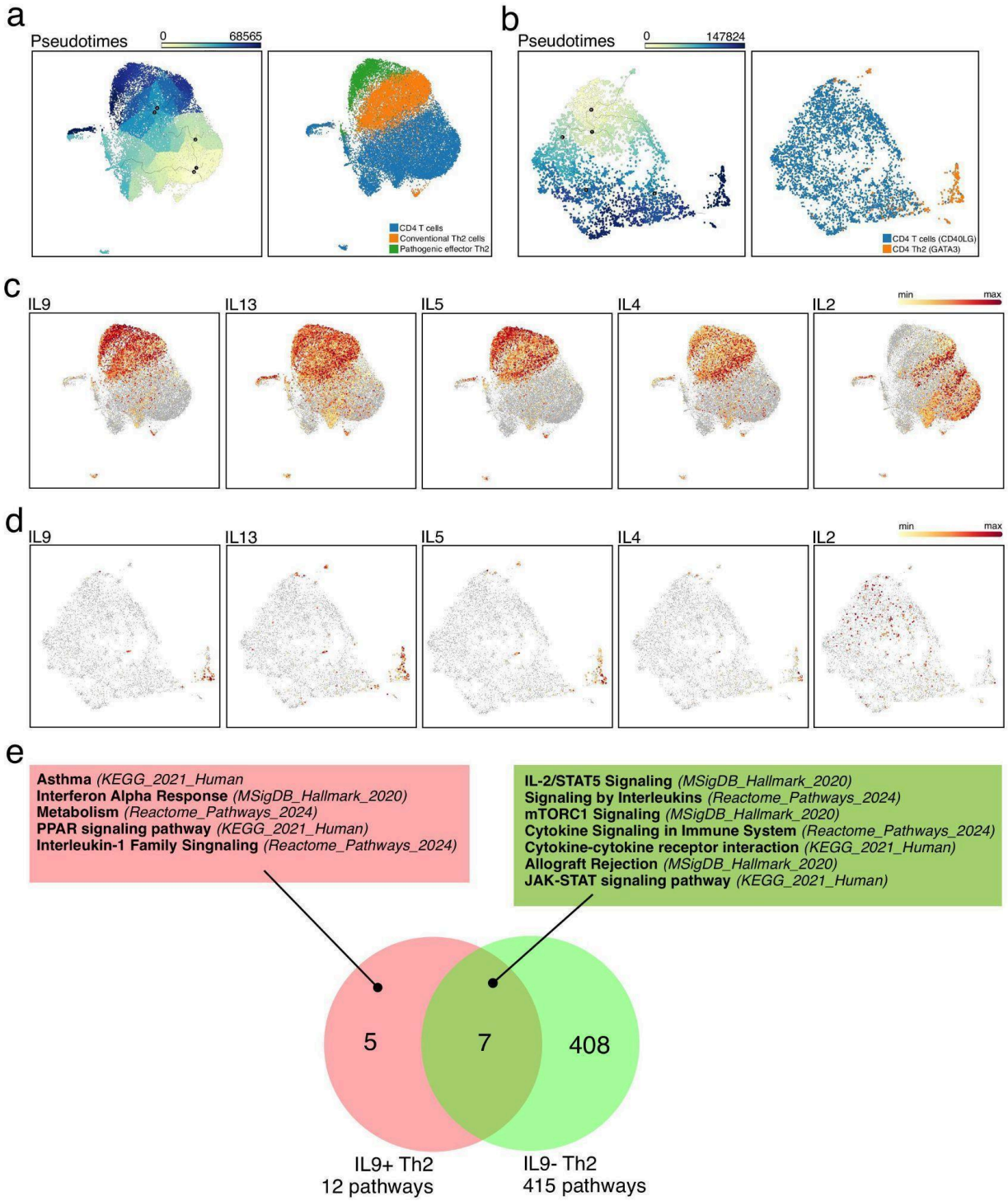

a
Pseudotimes
0
68565
CD4 T cells
Conventional Th2 cells
Pathogenic effector Th2
b
Pseudotimes
0
147824
CD4 T cells (CD40LG)
CD4 Th2 (GATA3)
c
IL9
IL13
IL5
IL4
IL2
min
max
d
IL9
IL13
IL5
IL4
IL2
min
max
e
Asthma (KEGG_2021_Human
Interferon Alpha Response (MSigDB_Hallmark_2020)
Metabolism (Reactome_Pathways_2024)
PPAR signaling pathway (KEGG_2021_Human)
Interleukin-1 Family Singnaling (Reactome_Pathways_2024)
IL-2/STAT5 Signaling (MSigDB_Hallmark_2020)
Signaling by Interleukins (Reactome_Pathways_2024)
mTORC1 Signaling (MSigDB_Hallmark_2020)
Cytokine Signaling in Immune System (Reactome_Pathways_2024)
Cytokine-cytokine receptor interaction (KEGG_2021_Human)
Allograft Rejection (MSigDB_Hallmark_2020)
JAK-STAT signaling pathway (KEGG_2021_Human)
5
7
408
IL9+ Th2
12 pathways
IL9- Th2
415 pathways

Figure 7. Pseudotime Trajectory and Pathway Enrichment Analysis of the IL9+ CD4+Th2 cell Subpopulation in EE and Asthma

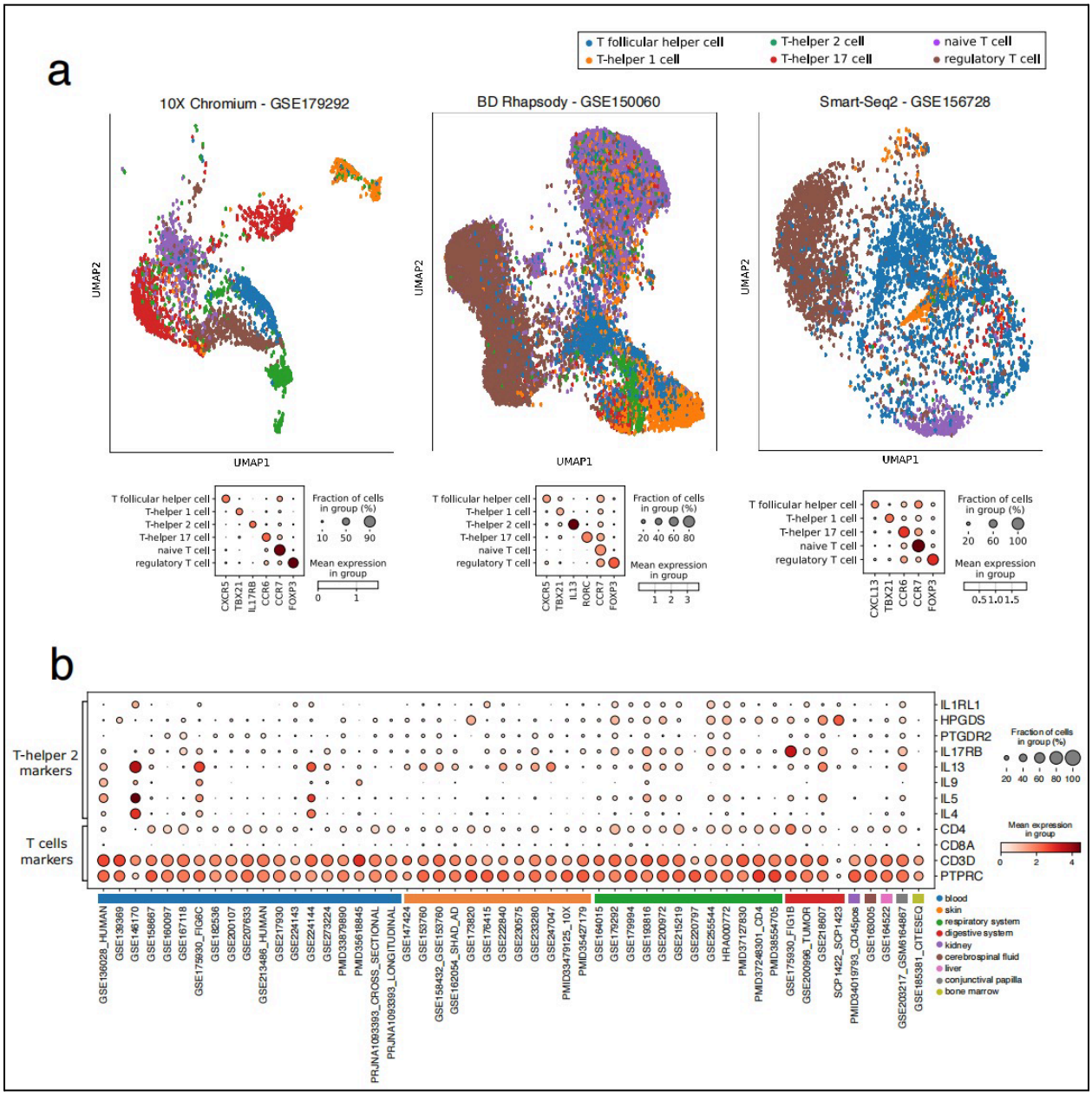
a-d. Pseudotime trajectory analysis of CD4 T cells, including CD4+Th2 cells, from **(a)** EE (GSE175930, n = 30,635 cells) and **(b)** asthma (GSE193816, n = 4,503 cells) datasets, with corresponding expression of *IL9, IL13, IL5, IL4,* and *IL2* shown in **(c)** and **(d)**, respectively.

c, d. Given that *IL2* and *IL4* are typically associated with the early stages of CD4+Th2 polarization and activation, early production of IL2 was observed in the EE blood datasets.

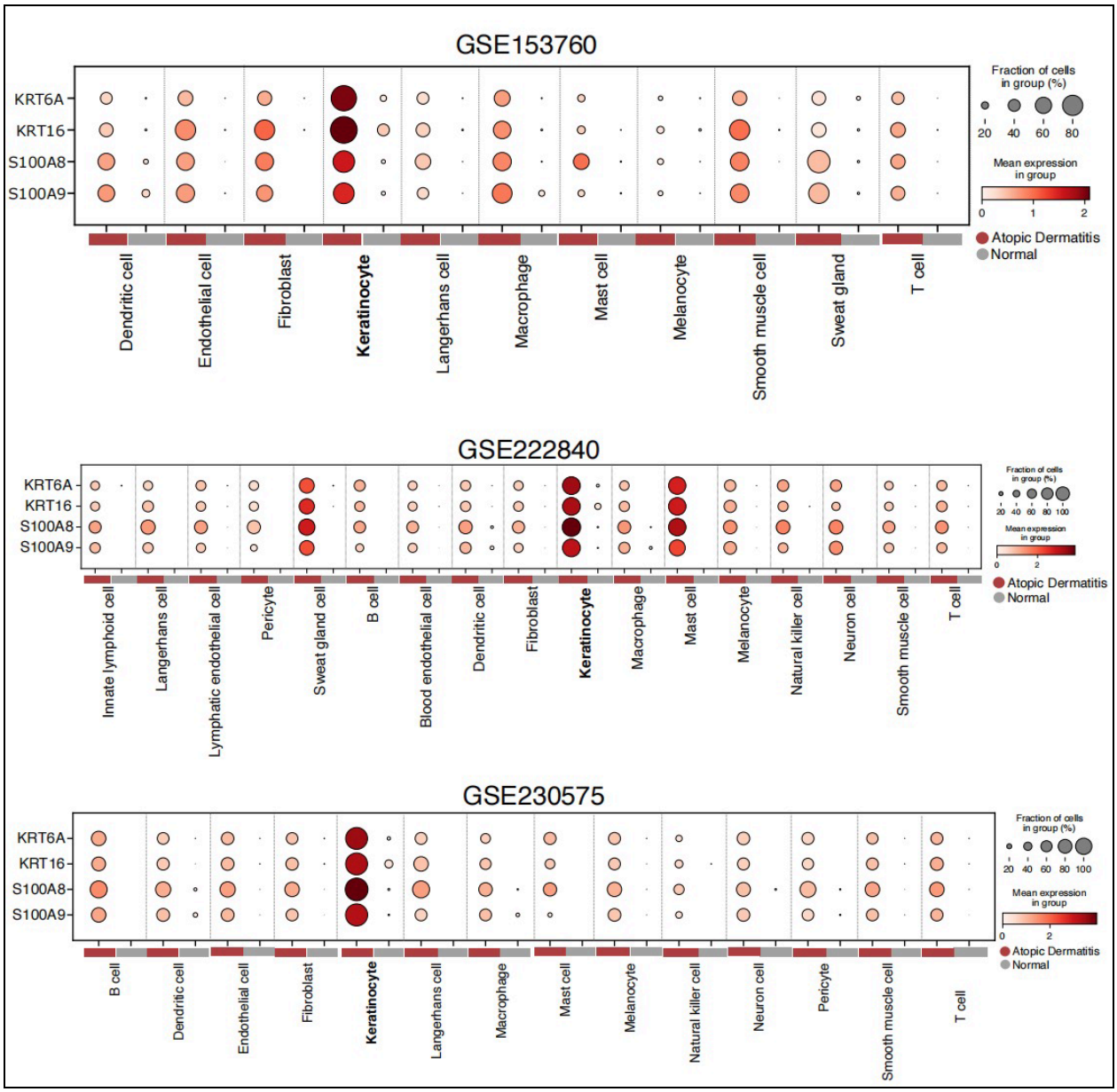
d. Th2 cells reverting from IL9 positive back to IL9 negative expression at the pseudotime endpoint

e. Pathway enrichment analysis comparing *IL9*-positive (IL9+ Th2) and *IL9*-negative (IL9- Th2) Th2 cells. The Venn diagram highlights the overlap of significantly enriched pathways, with 5 pathways unique to *IL9*+ CD4+Th2 cells, 408 pathways specific to *IL9*- CD4+Th2 cells, and 7 pathways shared between the two populations. Key enriched pathways are annotated for each category.

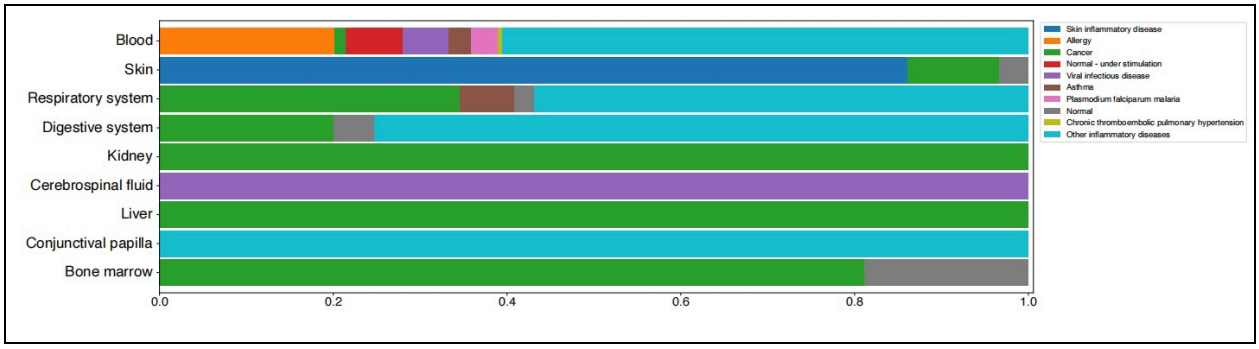
SUPPLEMENTARY DATA



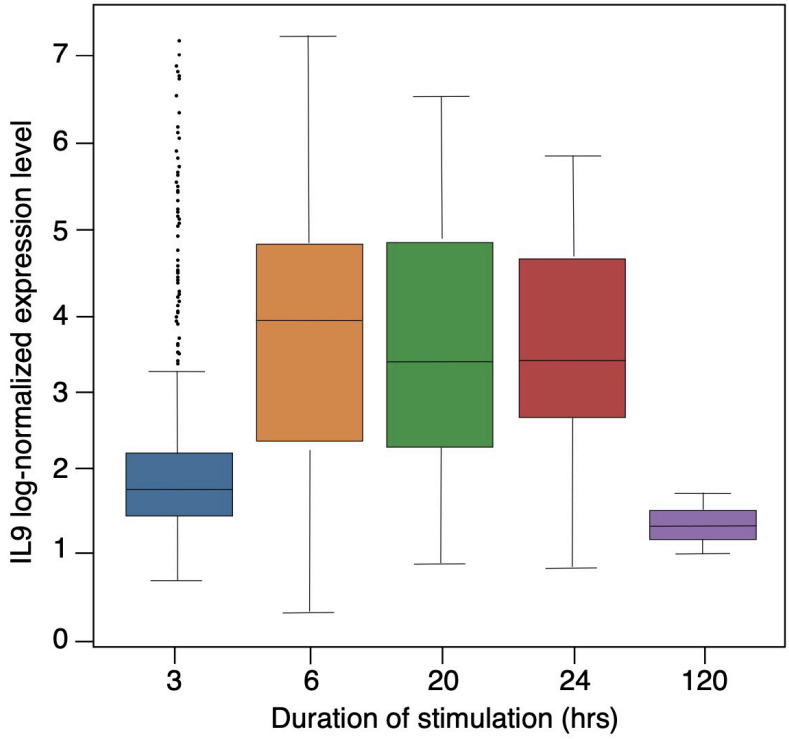
Supplementary Figure 1. CD4+ Th2 cell type prediction results.
a. CD4+Th2 cell type prediction results on different sequencing technologies.
b. CD4+Th2 cell marker expression across individual studies included in the CD4+ Th2 cell atlas, grouping by tissue of origin.



Supplementary Figure 2. Comparison of KRT6A, KRT16, S100A8, and S100A9 expression between AD and healthy samples. The comparisons were run on 3 individual studies and on each cell type separately. All cell types showed consistently higher expression of all 4 genes in AD compared to healthy samples, with the highest significance in keratinocytes.



Supplementary Figure 3. Proportions of conditions within each tissue



Supplementary Figure 4. Box plot of IL9 normalized expression across different allergen contact/stimulation durations. Across different allergic conditions in the current CD4+Th2 atlas. 3hrs: Food allergy (blue; GSE136028_HUMAN); 6hrs: EE and A/ARD (orange; GSE175930, GSE146170); 20hrs: Food allergy (green; GSE158667); 24hrs: A/ARD (red, GSE193816, GSE164059); 120hrs: allergic rhinitis (purple; GSE200107).

Supplementary Table 1:

Datasets included in the CD4+ Th2 Atlas

Th2 datasets

References:

1. Zhu, J. & Paul, W. E. Heterogeneity and plasticity of T helper cells. *Cell Res.* **20**, 4–12 (2010).
2. Moreno, A. S., McPhee, R., Arruda, L. K. & Howell, M. D. Targeting the T Helper 2 Inflammatory Axis in Atopic Dermatitis. *Int. Arch. Allergy Immunol.* **171**, 71–80 (2016).
3. Harker, J. A. & Lloyd, C. M. T helper 2 cells in asthma. *J. Exp. Med.* **220**, e20221094 (2023).
4. Hamilton, J. D. *et al.* Dupilumab suppresses type 2 inflammatory biomarkers across multiple atopic, allergic diseases. *Clin. Exp. Allergy* **51**, 915–931 (2021).
5. Sun, L., Su, Y., Jiao, A. et al. T cells in health and disease. *Sig Transduct Target Ther* **8**, 235 (2023)
6. Glatman Zaretsky, A. *et al.* T follicular helper cells differentiate from Th2 cells in response to helminth antigens. *J. Exp. Med.* **206**, 991–999 (2009).
7. Xu, L., Kitani, A., Fuss, I. & Strober, W. Cutting edge: regulatory T cells induce CD4+CD25-Foxp3- T cells or are self-induced to become Th17 cells in the absence of exogenous TGF-beta. *J. Immunol. Baltim. Md 1950* **178**, 6725–6729 (2007).
8. Sabat, R., Wolk, K., Loyal, L., Döcke, WD., Ghoreschi, K. T cell pathology in skin inflammation. *Semin Immunopathol.* **41**, 359-377 (2019).
9. BioTuring Talk2Data. https://app.bioturing.com/talk2data/.
10. Mitson-Salazar, A. & Prussin, C. Pathogenic Effector Th2 Cells in Allergic Eosinophilic Inflammatory Disease. *Front. Med.* **4**, 165 (2017).
11. Nakayama, T. *et al.* Th2 Cells in Health and Disease. *Annu. Rev. Immunol.* **35**, 53–84 (2017).
12. Kokubo, K. *et al.* Conventional and pathogenic Th2 cells in inflammation, tissue repair, and fibrosis. *Front. Immunol.* **13**, 945063 (2022).
13. Wang, S. *et al.* S100A8/A9 in Inflammation. *Front. Immunol.* **9**, (2018).

14. Boniface, K. *et al.* IL-22 inhibits epidermal differentiation and induces proinflammatory gene expression and migration of human keratinocytes. *J. Immunol. Baltim. Md 1950* **174**, 3695–3702 (2005).
15. Brunner, P. M. *et al.* Baseline IL-22 expression in patients with atopic dermatitis stratifies tissue responses to fezakinumab. *J. Allergy Clin. Immunol.* **143**, 142–154 (2019).
16. Bangert, C. *et al.* Persistence of mature dendritic cells, TH2A, and Tc2 cells characterize clinically resolved atopic dermatitis under IL-4Rα blockade. *Sci. Immunol.* **6**, eabe2749 (2021).
17. Alladina, J. *et al.* A human model of asthma exacerbation reveals transcriptional programs and cell circuits specific to allergic asthma. *Sci. Immunol.* **8**, eabq6352 (2023).
18. Semple, R. K., Chatterjee, V. K. K. & O'Rahilly, S. PPAR gamma and human metabolic disease. *J. Clin. Invest.* **116**, 581–589 (2006).
19. Furue, M. Regulation of Skin Barrier Function via Competition between AHR Axis versus IL-13/IL-4–JAK–STAT6/STAT3 Axis: Pathogenic and Therapeutic Implications in Atopic Dermatitis. *J. Clin. Med.* **9**, 3741 (2020).
20. Micossé, C. *et al.* Human "TH9" cells are a subpopulation of PPAR-γ+ TH2 cells. *Sci. Immunol.* **4**, eaat5943 (2019).
21. Schärli, S. *et al.* IL-9 sensitizes human TH2 cells to proinflammatory IL-18 signals in atopic dermatitis. *J. Allergy Clin. Immunol.* **155**, 491-504.e9 (2025).
22. Hauber, H.-P., Bergeron, C. & Hamid, Q. IL-9 in allergic inflammation. *Int. Arch. Allergy Immunol.* **134**, 79–87 (2004).
23. Zhao, H. *et al.* IL-9 neutralizing antibody suppresses allergic inflammation in ovalbumin-induced allergic rhinitis mouse model. *Front. Pharmacol.* **13**, (2022).
24. Bertschi, N. L. *et al.* PPAR-γ regulates the effector function of human T helper 9 cells by promoting glycolysis. *Nat. Commun.* **14**, 2471 (2023).
25. Vuong, H., Truong, T., Phan, T., Pham, S. Venice: A New Algorithm for Finding Marker

Genes in Single-Cell Transcriptomic Data. *bioRxiv.* (2020).

26. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
27. Mosmann, T. R. & Coffman, R. L. TH1 and TH2 Cells: Different Patterns of Lymphokine Secretion Lead to Different Functional Properties. *Annu. Rev. Immunol.* **7**, 145–173 (1989).
28. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
29. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M. k-Nearest Neighbor Classification. in *Data Mining in Agriculture* (eds. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M.) 83–106 (2009).
30. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
31. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
32. Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).
33. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
34. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).
35. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
36. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
37. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
38. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29

(2000).

39. The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

40. Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B. and Petryszak, R. The reactome pathway knowledgebase 2024. *Nucleic acids research*, **52**(D1), D672-D678 (2024).