Interpretable Inflammation Landscape of Circulating Immune cells

Laura Jiménez-Gracia[1,*], Davide Maspero[1,*], Sergio Aguilar-Fernández[1,*], Francesco Craighero[2,*], Sara Ruiz[1], Domenica Marchese[1], Ginevra Caratù[1], Marc Elosua-Bayes[1], Mohamed Abdalfatah[1], Angela Sanzo-Machuca[3,4], Ana M. Corraliza[3,4], Ramon Massoni-Badosa[1], Hoang A. Tran[5,6,7], Rachelly Normand[5,6,7], Jacquelyn Nestor[5,6,7], Yourae Hong[8], Tessa Kole[9,10], Petra van der Velde[9,11], Frederique Alleblas[9,11], Flaminia Pedretti[12], Adrià Aterido[13,14], Martin Banchero[9,11], German Soriano[15,16], Eva Román[16,16], Maarten van den Berge[9,10], Azucena Salas[3,4], Jose Manuel Carrascosa[17], Antonio Fernández Nebro[18], Eugeni Domènech[19], Juan Cañete[20], Jesús Tornero[21], Javier Pérez-Gisbert[22], Ernest Choy[23], Giampiero Girolomoni[24], Britta Siegmund[25], Antonio Julià[13,14], Violeta Serra[12], Roberto Elosua[26,27,28], Sabine Tejpar[8], Silvia Vidal[29], Martijn C. Nawijn[9,11], Sara Marsal[13,14,30], Pierre Vandergheynst[2], Alexandra-Chloé Villani[5,6,7], Juan C. Nieto[1,†], Holger Heyn[1,†]

Affiliations

1 Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain.
2 Signal Processing Laboratory 2 (LTS2), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
3 Inflammatory Bowel Disease Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
4 Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Spain.
5 Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, 02129, Massachusetts, USA.
6 Broad Institute of MIT and Harvard, Cambridge, 02142, Massachusetts, USA.
7 Harvard Medical School, Boston, Massachusetts, USA.
8 Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium.
9 Groningen Research Institute for Asthma and COPD (GRIAC), University Medical Center Groningen, Groningen, Netherlands.
10 Department of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.
11 Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.
12 Experimental Therapeutics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain.
13 Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain.
14 IMIDomics, Inc.
15 Department of Gastroenterology, Biomedical Research Institut Sant Pau (IIB Sant Pau), Barcelona, Spain.
16 Biomedical Research Network on Hepatic and Digestive Diseases (CIBEREHD), Instituto de Salud Carlos III. Madrid, Spain.
17 Dermatology Department, Hospital Universitari Germans Trias i Pujol. Badalona, Spain.
18 Rheumatology Department, Hospital Regional Universitario Carlos Haya. Málaga, Spain.
19 Gastroenterology Department, Hospital Universitari Germans Trias i Pujol, Badalona, Spain.
20 Rheumatology Department, Fundació Clínic per a la Recerca Biomèdica. Barcelona, Spain.
21 Rheumatology Department, Hospital Universitario Guadalajara. Guadalajara, Spain.
22 Gastroenterology Department, Hospital Universitario de la Princesa. Madrid, Spain.
23 Section of Rheumatology, Cardiff University, Cardiff, United Kingdom
24 Section of Dermatology and Venereology, University of Verona, 37129 Verona, Italy.
25 Department of Gastroenterology, Rheumatology and Infectious Diseases, Charité-Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany.
26 Hospital del Mar Research Institute (IMIM). Barcelona, Catalonia, Spain.
27 CIBERCV, Instituto de Salud Carlos III. Madrid, Spain.
28 Faculty of Medicine, University of Vic-Central University of Catalonia. Vic, Catalonia, Spain
29 Group of Immunology-Inflammatory Diseases, Biomedical Research Institut Sant Pau (IIB Sant Pau), Barcelona, Spain.
30 On behalf of IMID-Consortium.

* These authors contributed equally
† Joint senior authors.
To whom correspondence should be addressed.
Juan C. Nieto (juan.nieto@cnag.eu);
Holger Heyn (holger.heyn@cnag.eu)

Abstract

Inflammation is a biological phenomenon involved in a wide variety of physiological and pathological processes. Although a controlled inflammatory response is beneficial for restoring homeostasis, it can become unfavorable if dysregulated. In recent years, major progress has been made in characterizing acute and chronic inflammation in specific diseases. However, a global, holistic understanding of inflammation is still elusive. This is particularly intriguing, considering the crucial function of inflammation for human health and its potential for modern medicine if fully deciphered. Here, we leverage advances in the field of single-cell genomics to delineate the full spectrum of circulating immune cell activation underlying inflammatory processes during infection, immune-mediated inflammatory diseases and cancer. Our single-cell atlas of >2 million peripheral blood mononuclear cells from 356 patients and 18 diseases allowed us to learn a foundation model of inflammation in circulating immune cells. The atlas expanded our current knowledge of the biology of inflammation of acute (e.g. inflammatory bowel disease, sepsis) and chronic (e.g. cirrhosis, asthma, and chronic obstructive pulmonary disease) disease processes and laid the foundation to develop a precision medicine framework using unsupervised as well as explainable machine learning. Beyond a disease-centered classification, we charted altered activity of inflammatory molecules in peripheral blood cells, depicting functional biomarkers to further understand mechanisms of inflammation. Finally, we have laid the groundwork for developing precision medicine diagnostic tools for patients experiencing severe acute or chronic inflammation by learning a classifier for inflammatory diseases, presenting cells in circulation as a powerful resource for patient stratification.

Introduction

Inflammation is a biological response or state of the immune system that serves to protect the human body from environmental challenges, thereby preserving homeostasis and structural integrity of tissues and organs[1]. Inflammatory processes are activated in response to various triggers, such as infection or injury, and involve a multistep defensive mechanism aimed at eliminating the source of perturbation[2–4]. Thus, inflammation represents an altered state within the immune system, which can manifest as either a protective or pathological response[5]. The cellular and molecular mediators of inflammation play pivotal roles in nearly every human disease, encompassing a wide array of biological processes, including the complex interplay of cytokines, myeloid and lymphoid cells[6].

The initiation of inflammatory processes is driven by cellular stimulation, triggered by the release of proinflammatory cytokines[7,8]. These cytokines exert autocrine and paracrine effects, activating endothelial cells, subsequently increasing vascular permeability. This allows immune cells to infiltrate tissues at the site of infection, facilitated by chemokines. Chemokines are essential for recruiting additional immune cells, playing a crucial role in phagocytosis and pathogen eradication[9]. In the bloodstream, activated immune cells release cytokines and travel to various tissues. Inflammation is a central driver in cardio-vascular[10], autoimmune[11,12], infectious diseases[13,14] and even cancer[15]. The success of therapies targeting inflammation underscores the importance of understanding the underlying pathways[16–18]. Thus, categorizing patients based on their specific inflammatory cell states in the bloodstream has significant potential for advancing disease management[19].

Single-cell RNA sequencing (scRNA-seq) is becoming a conventional method for detecting altered cell states in blood, enabling the comparison of transcriptional profiles during perturbations, including inflammation[20]. Previous works revealed cellular profiles across diverse conditions, creating a shared phenotypic space that facilitates comparisons among patients and conditions, and generating a comprehensive view of inflammation[21]. Consequently, differential analysis of cell states and gene expression programs can now guide a holistic understanding of inflammation in acute and chronic diseases to form the basis for future precision medicine tools in diagnostics and novel treatments. In this regard, interpretable machine learning will play a pivotal role to extract disease-driving features from large healthy and disease single-cell references. Eventually, comprehensive models will allow the classification of patients for precise diagnostics and the patient stratification for tailored treatments.

Our study initially defined common immune cell types in peripheral blood, before capturing disease-specific inflammatory cell states that exhibit functional specialization within the inflammatory landscape. Beyond a disease-centered classification, we modeled the expression profiles of inflammatory molecules to define interpretable biomarkers driving immune cell activation, migration, cytotoxic responses, and antigen presentation activities. Ultimately, we developed a classifier based on the peripheral blood mononuclear cell (PBMC) reference, establishing inflammatory immune cell features as a precision medicine diagnostic tool for patients suffering from severe acute or chronic inflammation.

Main

An inflammation landscape of circulating immune cells

To chart a comprehensive landscape of immune cells in circulation of healthy individuals and patients suffering from inflammatory diseases, we analyzed the transcriptomic profiles of over 2 million (1.7 million after filtering) PBMCs, representing 356 patients and 18 diseases. Diseases broadly classified into five distinct groups: 1) Immune-mediated inflammatory diseases (IMIDs), 2) acute and 3) chronic inflammation, 4) infection and 5) cancers, which were profiled along with healthy donor samples (**Fig. 1a**). We completed our dataset (79% of total data) with additional studies to generate a comprehensive resource of immune cell states across inflammatory diseases and beyond (**Fig. 1a**; **Supplementary Table 1**). Our cohort included various scRNA-seq chemistries (10x Genomics 3' and 5' mRNA) and experimental designs (CellPlex and genotype multiplexing), as well as individuals of both sexes and across age groups, to comprehensively capture technical and biological variability (see **Methods**). To learn a generative model of circulating immune cells of inflammatory diseases, we applied probabilistic modeling of the single-cell data using scVI[22] and scGen[23], considering clinical characteristics (disease, sex and age, **Extended Data Fig. 1a,b**). scGen generates a lower-dimensional cell embedding space, before reconstructing gene expression data. Batch effects are removed based on gene-specific parameters, learned during the integration. Its generative probabilistic models proved superior performances in integrating complex datasets compared to other approaches, particularly if cell annotations are available (**Extended Data Fig. 1c-e**)[23]. Applied here, the resulting gene expression profiles and the cell embedding space were batch effect corrected, while preserving biological heterogeneity (i.e. previously annotated cell types and states; **Supplementary Table 2**). From the joint embedding space, we initially assigned cells to eight major cell types (Level 1; **Fig. 1b; Extended Data Fig. 1f**): (1) Lymphocytes B, (2) Lymphocytes T, (3) NK cells, (4) Monocytes, (5) Dendritic Cells (DC), (6) Hematopoietic Stem Cells (HSC), (7) Plasmacytoid Dendritic Cells (pDC), (8) Platelets, and Red Blood Cells (RBC). Following a recursive, top-down clustering approach (see **Methods**), we obtained a total of 69 subclusters (Level 2), comprehensively resembling immune cell states of the innate and adaptive compartments (**Extended Data Fig. 2**). Noteworthy, integrating a large number of patients and cells allowed a fine-grained description beyond previously annotated cell states (immune cell types with distinct activation-related transcriptomes; **Supplementary Table 2**).

Diving deeper into genes, programs and signatures to characterize inflammatory diseases, our subsequent analysis followed three complementary strategies to identify disease-driving mechanisms (gene signature activity), to define biomarkers for inflammatory responses (feature extraction) and to classify patients based on their disease-specific signatures (projection). Therefore, we looked at gene expression profiles holistically, but also delineated the inflammatory process by focusing on molecules that trigger immune cell activation, cellular migration and extravasation, antigen presentation and cytotoxic responses (**Supplementary Table 3**)[24–30]. These strategies jointly allowed us to enlarge our understanding of inflammatory processes and their contribution to inflammatory diseases, but also form the basis for precision medicine tools by establishing cells in circulation as potent biomarkers for disease diagnostics.

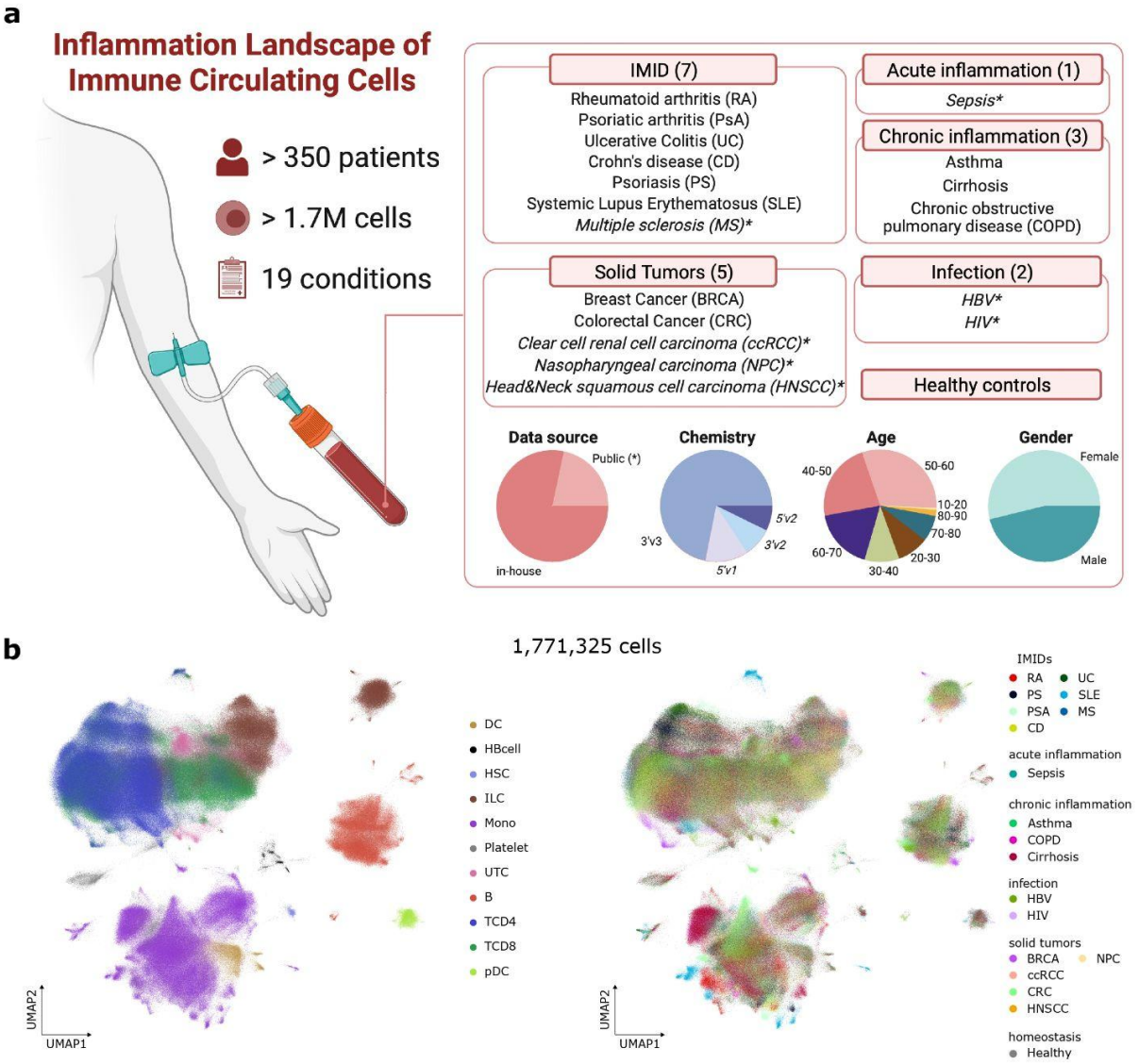

Figure 1. Inflammation Landscape of Circulating Immune Cells. (**a**) Schematic overview describing the number of cells, samples, conditions (diseases and disease groups) and patients with the associated metadata related to the origin (in-house, public data), scRNA-seq chemistry (10x Genomics assay and version) and patient (age and gender). * indicates public data. (**b**) Uniform manifold approximation and projection (UMAP) embedding for the scGen-corrected latent space considering the full dataset across patients and diseases (1,771,325 cells) colored by the major cell types and states (left, Level 1) and diseases (right).

Inflammation-related signatures across diseases and cell types

To identify inflammation-related signatures across cell types and diseases, we first ran a Multivariate Linear Model (MLM) analysis using DecoupleR[31] to assess the activation profiles of 12 inflammation-relevant gene set signatures (444 genes, **Supplementary Table 3**). The MLM was applied on the scGen-corrected dataset, providing an inflammation signature activity score for each cell, before averaging by disease and cell type (see **Methods**). Finally, we computed the relative difference between diseased and healthy samples to highlight disease-specific alterations.

Across all disease groups, we observe a general trend of increased activity in immune-relevant signatures as compared to healthy donors (>50% increased average signature scores; **Fig. 2a**). From all IMIDs, SLE showed a uniquely strong upregulation of the IFN response signature paired with a downregulation of chemokines receptors. Additionally, both SLE and MS showed a decreased anti-inflammatory cytokines receptor signature, whereas we observed an increase in all other IMIDs. All IMIDs, but MS, exhibited an upregulation of the Type I IFN signal, opposite to all other disease categories. As previously shown[32], we captured the upregulation of the TNF ligand signature for sepsis (together with an increase in antigen presentation molecules), with a decrease in the other inflammatory signals (chemokines and cytokines). In contrast, all chronic inflammatory diseases upregulated the activity of proinflammatory cytokine receptors, but showed decreased Type I IFN signaling. The Type I IFN signature was also decreased in viral infections, while we found an increased activity in all remaining inflammation-related pathways. Finally, within solid tumors, CRC and HNSCC presented a strong upregulation of TNF receptors.

Considering distinct cell types as unique contributors to the inflammatory immune landscape, we delineated signatures at cell type level (i.e. Level 1 and Level 2). In line with the aforementioned alterations, an elevated IFN response signature identified SLE patients as the strongest effector of IFN stimulation. Intriguingly, IFN response has been previously described to contribute to SLE through the activation of distinct immune cell types[33]. Here, we described a more systemic response with IFN response genes being activated in most major cell lineages (Level 1; **Extended Data Fig. 3a and 3b**). At immune subpopulation level (Level 2), the increased IFN response signature for SLE identified IGHG+ plasma cells, but also CD4, CD8 effector memory, inflammatory monocytes and dendritic cells (DC2B) as major contributors to the inflammatory process (**Fig. 2b and Extended Data Fig. 3b)**. Previous studies described the pathogenic effect of IFN in SLE through the overstimulation of abnormal germinal centers and the differentiation to pathogenic-associated plasmablasts and antigen presenting cells. Here, we showed a higher activity specifically in antibody secreting plasma cells (IGHG+), which have been related with the production of auto-antibodies in SLE[34]. The activation and organization of germinal centers depends on the stimulation of B cells by follicular helper CD4 T cells[35]. Interestingly, CD4 effector memory T cells in SLE presented an increased IFN response activity, but also expressed markers of T follicular helper cells, strongly suggesting their role in triggering abnormal T-B interaction and altered activation of B cells in germinal centers (**Fig. 2c)**. Activation and differentiation to T follicular helper cells further relies on antigen presenting cells, in line with the elevated IFN response detected in DC2B cells. We validated the results through the identification of gene expression factors across cell types using a data-driven approach (Spectra[36]), confirming IFN genes with correlating gene expression levels that differentiated IMIDs and specifically SLE from other diseases (**Extended Data Fig. 3c**).

Compared to most other diseases, the Type I IFN signature was upregulated in IMIDs (except MS), with particularly high activities in SLE and CD (**Fig 2a**). An excess of Type I IFN has been associated with the severity of systemic autoimmune diseases and auto-antibodies production[37] and with multiple effects on the adaptive immunity in CD[38]. At cell type and subpopulation level (Level 1 and Level 2), we observed an enrichment in almost all cell lineages (**Extended Data Fig. 3d**), with dendritic cell subtypes (pDC, DC2A and DC5) and non-classical monocytes showing the highest Type I IFN activity and

highlighting the role of the innate immunity in the systemic inflammation of CD (**Fig. 2d**). In line with previous observations[39], also Memory Switch and Memory ITGAX B cells showed an enriched Type I IFN signature (**Fig. 2d**). Particularly, ITGAX+ B cells are a rare B cell subtype with pathogenic activities in many autoimmune diseases, but are not well characterized in CD[40]. Previous work further pointed to the critical role of DCs for breaking peripheral tolerance to create a cytotoxic environment through the activation and recruitment of CD4 T-helper, NK and CD8 T in inflammatory bowel disease (IBD) and RA[38]. Exploring the landscape of inflammatory signatures in these populations, we observed an upregulation of chemokines, pro-inflammatory cytokines and TNF ligands in CD56dim NK activated, CD4 and CD8 effector memory, CD8 activated and CD8 IFN responder T cells (**Fig. 2e**). Such increased expression of major inflammatory signatures in the peripheral blood of CD patients, further strengthens the role of DC activation via Type I IFN to induce the lymphoid compartment and amplifying systemic inflammation in CD[41,42].

Functional biomarker selection through interpretable modeling

Biomarker discovery using linear models (such as the above applied MLM) or standard differential expression analysis suffers from the limitation that genes are considered independently. Thus, we considered the possibility of categorizing cells to their respective disease origin through an interpretable machine learning pipeline, to guide the selection of functional disease biomarkers (451 genes, **Supplementary Table 3**). Therefore, we next applied a supervised classification approach, together with a post-hoc interpretability method, to allow the inference of the gene-wise importance, stratified by disease. We based our strategy on Gradient Boosted Decision Trees (GBDTs), a state-of-the-art machine learning technique proven to be effective in complex tasks with noisy data and non-linear feature dependencies[43]. GBDTs iteratively build an ensemble of decision trees, by trading the complexity of the model (i.e. the number of trees) with the accuracy of the predictions. Here, we used the CatBoost library given its superior generalization performance[43] that required the definition of a set of hyperparameters, such as the maximum tree depth. To tune the hyperparameters, we employed the TPE (Tree-structured Parzen Estimator) sampling algorithm[44] implemented in the Optuna library[45] (see **Methods**). As GBDTs require post-hoc interpretability tools in order to infer explanations, we computed SHAP (SHapley Additive exPlanation) values[46], shown to provide attributions that are locally consistent and that can be aggregated into global explanations. SHAP values explained the output of the classifier, in our case the predicted disease, as a sum of contributions of each feature, that is, the gene profiles. Such contributions correspond to the change in the expected model prediction when conditioning on the considered feature.

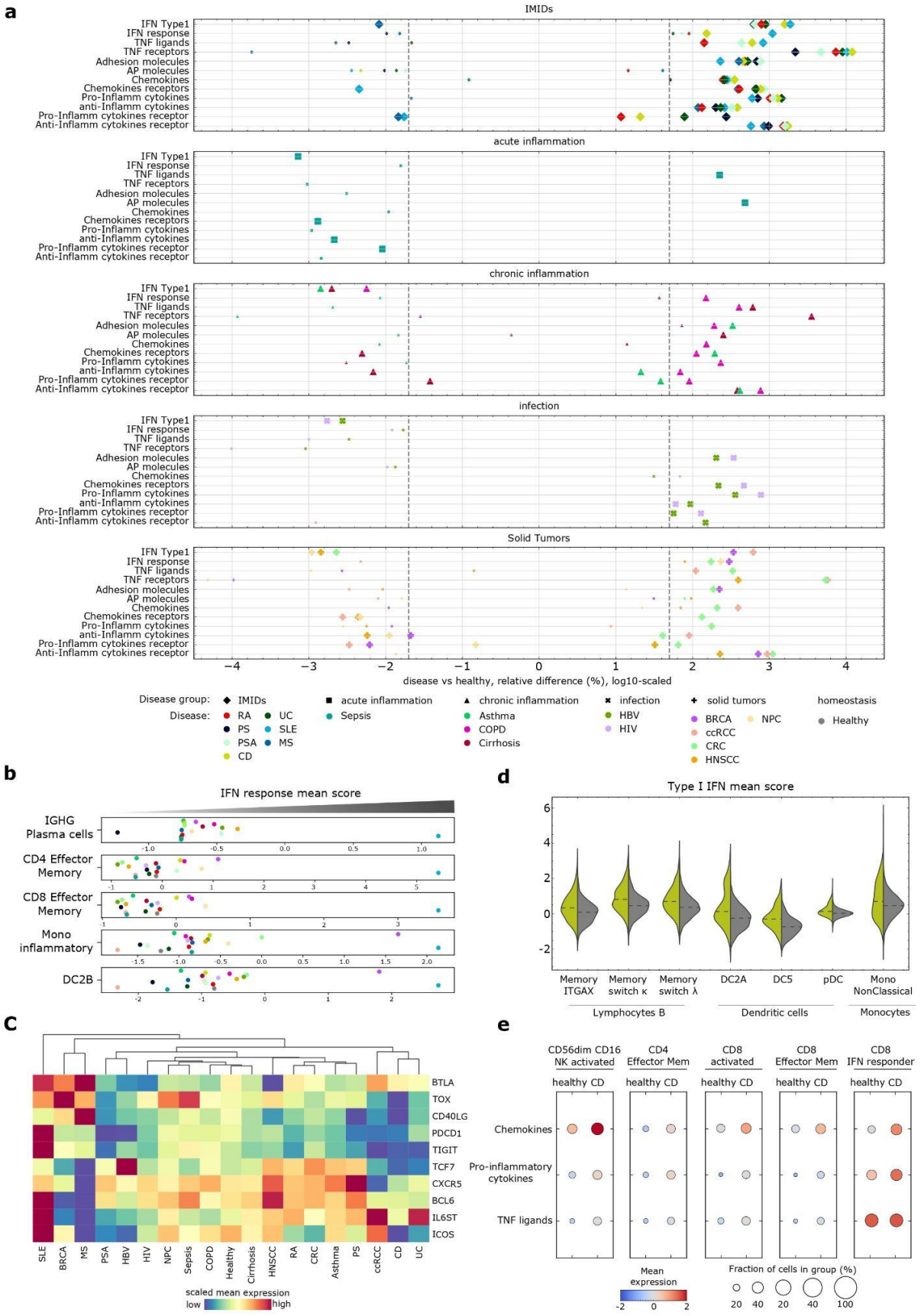
a
IMIDs
acute inflammation
chronic inflammation
infection
Solid Tumors
IFN Type1
IFN response
TNF ligands
TNF receptors
Adhesion molecules
AP molecules
Chemokines
Chemokines receptors
Pro-Inflamm cytokines
anti-Inflamm cytokines
Pro-Inflamm cytokines receptor
Anti-Inflamm cytokines receptor
−4 −3 −2 −1 0 1 2 3 4
disease vs healthy, relative difference (%), log10-scaled
Disease group: IMIDs acute inflammation chronic inflammation infection solid tumors homeostasis
Disease: RA UC PS SLE PSA MS CD Sepsis Asthma COPD Cirrhosis HBV HIV BRCA ccRCC CRC HNSCC NPC Healthy
b
IFN response mean score
IGHG Plasma cells
CD4 Effector Memory
CD8 Effector Memory
Mono inflammatory
DC2B
d
Type I IFN mean score
Memory ITGAX
Memory switch κ
Memory switch λ
DC2A
DC5
pDC
Mono NonClassical
Lymphocytes B
Dendritic cells
Monocytes
c
BTLA
TOX
CD40LG
PDCD1
TIGIT
TCF7
CXCR5
BCL6
IL6ST
ICOS
SLE BRCA MS PSA HBV HIV NPC Sepsis COPD Healthy Cirrhosis HNSCC RA CRC Asthma PS ccRCC CD UC
scaled mean expression
low high
e
CD56dim CD16 NK activated
CD4 Effector Mem
CD8 activated
CD8 Effector Mem
CD8 IFN responder
healthy CD
Chemokines
Pro-inflammatory cytokines
TNF ligands
Mean expression
-2 0 2
Fraction of cells in group (%)
20 40 20 40 100

Figure 2. Inflammation profiles across cell types and diseases. (**a**) Scatterplot displaying relative difference (log-10 scaled) across diseases (colors) versus healthy controls, categorized by disease groups (symbols) for 12 immune-related signature pathways. Positive signature activity score (large icons) and significant changes (>50% relative increase, dashed lines) are indicated. (**b**) Scatterplot showing the mean IFN response signature score across selected cell types (level 2) and colored by disease (as in **a**). (**c**) Heatmap presenting the scGen-corrected mean expression of follicular peripheral T cells (Tph) genes across diseases (data scaled by gene). (**d**) Violinplot displaying the Type I IFN signature score distribution across selected cell types (level 2) comparing CD patients and healthy donors. (**e**) Dotplot with the mean chemokine, pro-inflammatory cytokine and TNF ligand signature scores (y-axis) at cell types (level 2; x-axis) in CD patients and healthy donors. The dot size reflects the percentage of cells in a cluster (from where the gene signature was computed) and the color represents the average score levels.

Combining the two approaches, we were able to rank genes based on their importance stratified by cell type and disease. We executed the analysis considering each cell type (Level 1) independently, resulting in gene rankings based on importance to classify diseases. Such strategy mitigated the impact of cell-specific expression profiles and allowed the interrogation of differential gene importance to distinguish diseases in distinct cell types. We applied the pipeline on the corrected gene expression profiles after scGen integration, but also tested uncorrected log-normalized data as an input. Overall, we achieved high accuracy to assign each cell to the correct disease label (balanced accuracy score = 0.78, computed on a test set of 20% cells; **Fig. 3a**), when starting from scGEN-corrected data. Instead, log-normalized counts resulted in decreased balanced accuracy scores proving the improvement provided after batch correction (0.63; **Fig. 3b**). Noteworthy, HIV and sepsis obtained lower accuracy scores (0.33 and 0.5, respectively), likely due to the low cell numbers (~0.5% of total cells) and notable differences across cell types (**Extended Data Fig. 4**).

Leveraging the interpretable machine learning pipeline and the well-annotated gene sets for biomarker discovery resulted in a rich resource of prioritized biomarker genes and their related cell types. Ordering genes by their importance within immune cell types, *CYBA* stood out as a strong candidate marker gene for IMIDs affecting barrier tissues, particularly Crohn's disease (CD), Ulcerative colitis (UC), Psoriasis (PS) and Psoriatic Arthritis (PSA) (**Fig. 3c**). Interestingly, elevated levels of *CYBA* were important for intestinal inflammatory diseases (CD, UC), whereas reduced levels were related to diseases manifesting in the skin (PS, PSA, **Fig. 3d**). *CYBA* encodes the primary component of the microbicidal oxidase system of phagocytes. In line, the importance of the gene was seen only in myeloid cells, particularly Monocytes and DCs (**Fig. 3d**). Mutations in *CYBA* cause an autosomal recessive chronic granulomatous disease and patients show an impaired phagocyte activation and fail to generate superoxide. Consequently, patients show recurrent bacterial and fungal infections in barrier tissues, including the skin[47]. Thus, we hypothesize that reduction of *CYBA* in skin-related IMIDs leads to an impaired immune barrier function and frequent recurrent infections causing localized, symptomatic flares of PS and PSA.

In CD and UC, both subtypes of IBD, upregulation of *CYBA* may result in the accumulation of Reactive Oxidative Species (ROS), a hallmark of both diseases. ROS produced by mucosa-resident cells or by newly recruited innate immune cells are essential for antimicrobial mucosal immune responses and defense against pathogenic attack[48]. We confirmed the overexpression of *CYBA* in myeloid cells of UC patients and healthy donors in an independent PBMC validation cohort[49] ($p<0.01$; **Fig. 3e**) and within

intestinal tissues of CD, UC and healthy control donors (scRNA-seq, n=18; **Fig. 3f,g**)[50]. In the tissue biopsies, especially monocyte-derived (M0) and tissue-resident (M2) phagocytic macrophage populations showed significant upregulation of *CYBA* gene expression levels compared to healthy controls (p<0.01).

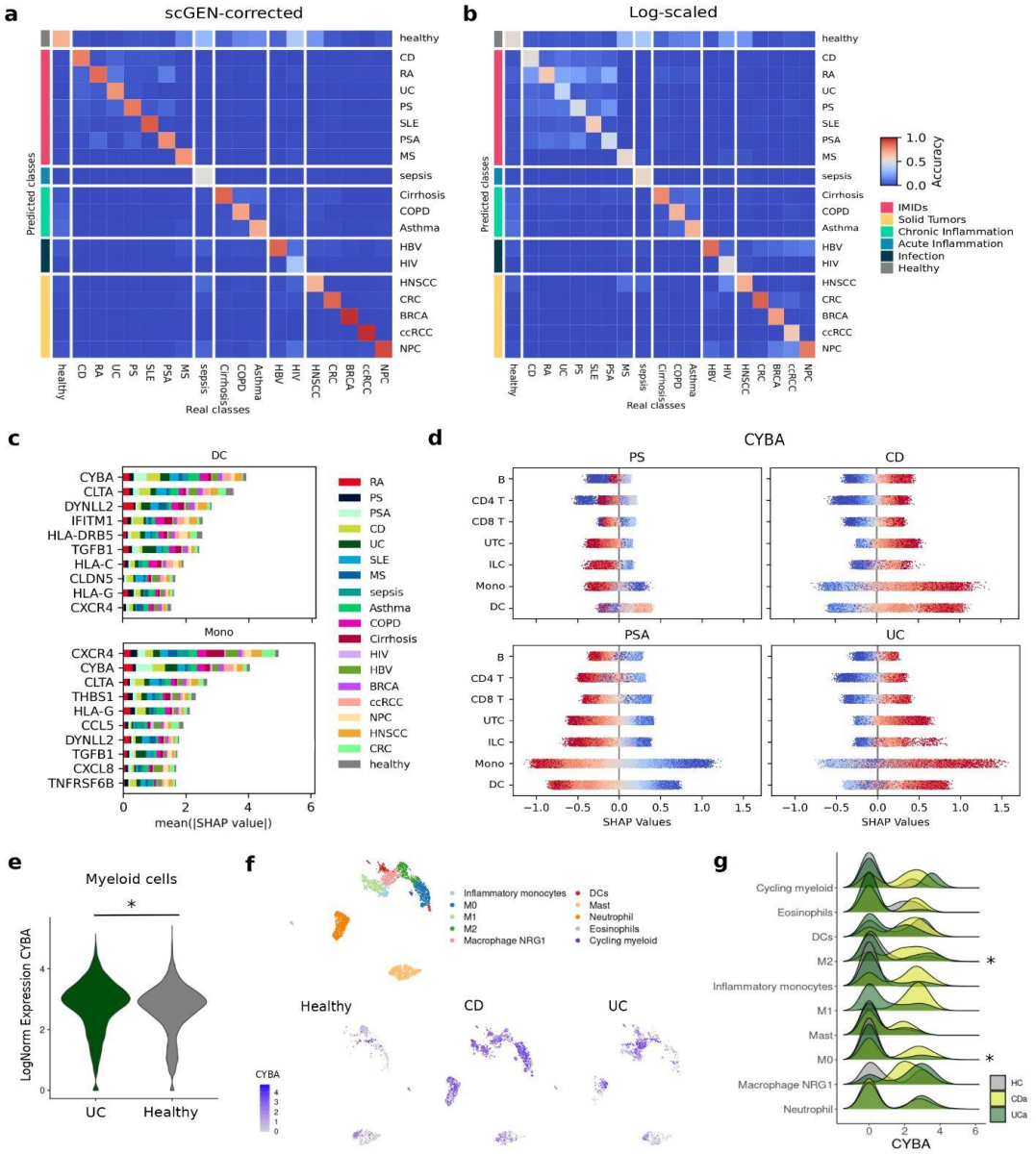

Fig. 3. Biomarker discovery using interpretable machine learning. (**a,b**) Diffusion matrices displaying the patient stratification accuracy to assign each cell to the correct disease label using the scGen-corrected (a) or uncorrected log-scaled data (b) as an input. (**c**) Gene list ranked by importance (mean absolute SHAP value) to classify diseases stratified by cell type (DC and Monocytes; level 1). (**d**) each row includes 10000 randomly sampled SHAP values, displaying the importance of *CYBA* gene expression as biomarker for PS, PSA, CD and UC across cell types (red: expression gain; blue: expression loss). (**e**) *CYBA* expression levels in an independent single-cell PBMC validation cohort[49] of UC and healthy donors. (**f**) Single-cell analysis from CD, UC and healthy patient biopsies; uniform manifold approximation and projection (UMAP) of myeloid cells coloured by cell type (top) and *CYBA* expression levels (bottom), using a blue color scale, across disease conditions. (**g**) *CYBA* expression levels in CD, UC and healthy patient biopsies stratified by myeloid subpopulations. Asterisks (*) indicates statistical significant changes using the Wilcoxon signed-rank test.

Ranking genes by their importance across diseases, *IFITM1* stood out as a biomarker for COPD (**Extended Data Fig. 5a**). In contrast to *CYBA*, the importance of *IFITM1* was mainly observed in lymphoid cells (CD4 T, CD8 T and ILC cells; **Extended Data Fig. 5b**). In line, *IFITM1* expression in these cells was higher in the blood of COPD patients compared to healthy controls (**Extended Data Fig. 5c**). *IFITM1* has an immune-modulatory effect controlling proliferation, adhesion and migration of CD4 T, CD8 T and ILC[51]. ILC and CD8 T cell accumulation is associated with decline of lung function and severity in COPD patients, and CD4 T cells mediate autoimmune response in COPD facilitating B-cell production of IgG autoantibodies in those patients[52]. We hypothesize the higher expression of *IFITM1* in lymphoid cells to be a mechanism of the accumulation of lymphoid cells induced by chronic inflammation[53].

Classification by patient projection into the embedding space

The ability to accurately classify different cell types according to their respective diseases prompted us to classify patients based on their disease of origin, creating the basis for a universal classifier as a precision medicine tool for inflammatory disease prediction. Single-cell information layed a foundation for better understanding the diversity and traits within the populations, but classifying new patients remains a challenge due to data sparsity and noise. By considering each patient as an ensemble of expression profiles across all cells, we learned a generative model during cell integration as a basis to project new patients into the same embedding space.

Projecting expression data into a lower dimensional space is a common strategy to reduce noise[54]. Here, we propose a novel computational framework to exploit the cell embedding for classification of patients into conditions (e.g. inflammatory diseases), thus, turning the single-cell reference into a diagnostic tool (**Fig. 4a**). Therefore, we first generated a cell type pseudobulk profile per patient by averaging the embedded features of the corresponding cells (annotation Level 1; see **Methods**). Next, we trained an independent linear classifier to assign correct disease labels, considering one cell type at a time. We handled uncertainty at cell type level via majority-voting system to determine most frequent conditions (see **Methods**). To assess the accuracy of our framework, we implemented a 5-fold cross validation strategy by splitting the full patient set into five balanced folds. The full pipeline was executed 5 times, considering each one fold as a test set and the remaining four as training sets. To further stress the ability of the approach to classify new patients, we removed cell annotations in the test set, before transferring the annotation labels from the training set after learning the embedding space[55,56].

Strikingly, our classification strategy resulted in a balanced accuracy score (BAS), averaged across five independent runs, greater than 0.90±0.05 (minimum 0.82; **Fig. 4b**) and with very low False Negative rates (13/18 diseases classified with an accuracy >0.9). Intriguingly, even overall low cell label transfer accuracies (Level 1: 0.78±0.01 and Level 2: 0.5±0.01) allowed to correctly classify patients with highly balanced precision scores, further validating pseudobulk modeling as an effective strategy to reduce the noise during the single-cell data projection. Such high accuracy, strongly suggests gene expression profiles of inflammatory diseases to be separated in the embedded space and, more importantly, immune cells in circulation to be capable of serving as source for patient classification.

Training a classifier for each cell type separately allowed us to assess their relevance in distinguishing inflammatory diseases, particularly for diseases with a lower overall accuracy (**Fig. 4b,c**). While certain IMIDs (SLE, PS, PSA, and RA) were particularly well classified by lymphoid cell types (T, B and ILCs) with decent BAS scores also in myeloid cell types, HIV could only be classified by lymphoid cells (i.e., T-CD4, T-CD8, and B, 0.9±0.28), while myeloid cell types (i.e. Mono, DC, and pDC) did not allow correct disease assignment (0.31±0.4). Overall, pDC and DCs showed the lowest average accuracy (0.67±0.24 and 0.77±0.33), likely due to the fact that 22 out of 353 patients did not contribute pseudobulk profiles for these rare cell types, highlighting the strength of an integrated atlas and foundation model of inflammatory diseases to holistically model predictive features across cell types and diseases.

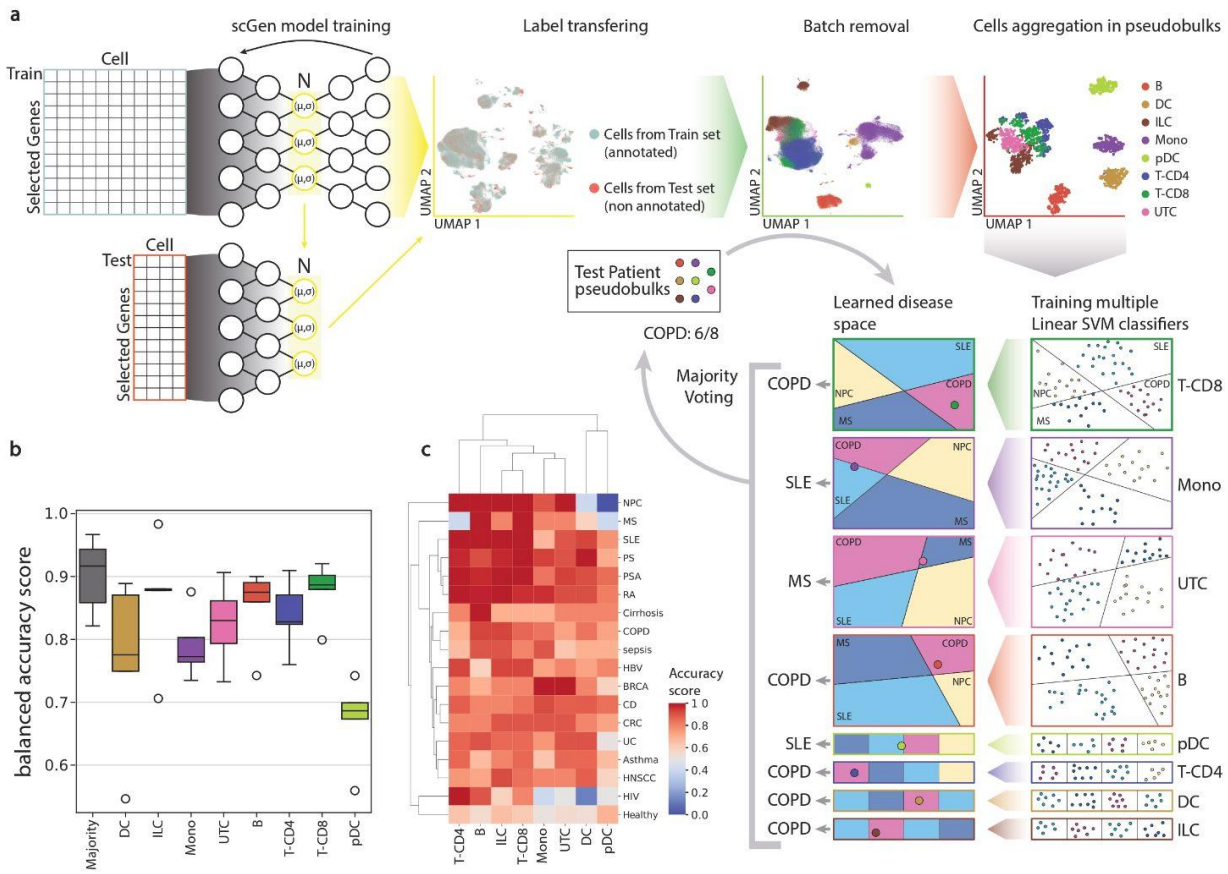

Figure 4. Schematic representation of the patient classifier framework. (**a**) First, the scGen foundation model is trained considering annotated datasets from all diseases. A common lower dimensional space is learned and exploited to transfer labels (Nearest Neighbor Descent via NNDescent) on the new patient samples after projecting the cell expression profiles into the same embedding. Second, technical confounding effects are removed using scGen to obtain a corrected cell embedding. Third, pseudobulk profiles are generated by averaging cell embedded features (n=200) grouped by cell types and patients. Fourth, one linear Support Vector Machine (SVM) classifier for each cell type is trained to assign the correct disease label to each patient. Finally, pseudobulks are generated for new patient samples and diseases are predicted using the linear classifiers. To aggregate the prediction into a unique disease label, a majority-voting approach was implemented. (**b**) Boxplot of balanced accuracy distribution across five independent runs, stratified by cell types. The 'Majority' label refers to the final label assignment. The full framework was executed five times removing 20% of patients from the known set and used to define the unknown set. (**c**) Heatmap showing the average accuracy stratified by cell types and diseases after the 5-cross validation step.

Discussion

Mapping the plasticity of the immune cells in circulation is now possible by using sequencing technologies that allow an unbiased immuno-phenotyping of single cells[57,58]. Importantly, such transcriptome-wide methods do not require previous knowledge, but leverage the random sampling of cells, and transcripts within cells, to derive comprehensive cellular landscapes. Recent technologies enable the sampling of thousands of cells per sample and hundred-thousands per patient cohort, pushing the resolution towards fine-grained cellular maps and increasing the power to identify disease-specific states and programs[59]. To date, single-cell sequencing has been applied to a multitude of inflammatory diseases to determine alterations in cell type composition and to pinpoint disease-driving mechanisms as potential therapeutic targets[20]. However, a complete map of immune cell states across diseases, holistically charting immune plasticity in inflammatory diseases, has been elusive. We reasoned that integrating single-cell transcriptome maps of cells in circulation across a variety of diseases and millions of cells would allow us to extract the full spectrum of features representing inflammatory processes and to generate a foundation model of inflammation in circulating immune cells.

Our strategy split the analysis into three phases, starting with the supervised extraction of inflammation-related patterns (gene signature activity), followed by the discovery of interpretable biomarkers (features) and finally patient classification (projection). Especially the latter presents an intriguing avenue towards personalized medicine strategies to manage and monitor inflammatory diseases and to move beyond current clinical practice using single biomarker strategies (e.g. CD4 T lymphocyte count in HIV)[60]. Here, the size of the datasets and the depth of cellular resolution, combined with an advanced machine learning framework, allowed the separation of diseases in the latent space and eventually the correct classification of unseen patient samples across disease categories and entities. Hence, we provide strong evidence for the potential of using immune cells in circulation as liquid biopsy for diagnostics, when combined with machine learning models that holistically learned the full spectrum of immune cell variability across patients and diseases[61]. Such models represent the cornerstone to build classifiers for inflammatory diseases as precision diagnostics tools for the clinic[62]. We note though that the here presented evidence provides the basis for establishing blood-based patient classification and stratification, but thorough follow-up work is required to determine the positive predictive value in even larger validation efforts.

However, the concept of using immune cells as a sensor for diseases is highly intriguing and opens the door for the development of future universal diagnostic tools[63]. For some of the tested diseases, such as sepsis, current biomarker strategies may already provide sufficient sensitivity considering the pathology and clinical manifestation of the disease (i.e. lymphopenia as a hallmark of sepsis)[32]. Nevertheless, for diseases such as in rheumatology and IBD, many patients are undiagnosed or diagnosed as False Positive, and more accurate universal tools are needed[64,65]. The here established classification framework was capable of classifying most diseases correctly using a majority-voting strategy across all cell types. However, for certain disease types, such as HIV or MS, a tailored, cell type-centric approach may lead to improved accuracy. Both diseases showed increased classification

scores using lymphoid cell types alone. In line, both represent T cell pathologies with MS being a T-cell–mediated autoimmune disease sustained by autoreactive T cells against myelin components and HIV infecting mainly CD4 T cells[66,67].

While the machine learning strategy developed for patient stratification was not interpretable, our complementary approach using GBDT, together with SHAP, and a curated list of immune cell molecules with defined function, provided explainable results and a rich resource for biomarker discovery[43,46]. Here, *CYBA*, the light, alpha subunit of microbicidal oxidases in phagocytes showed the strongest importance across all cell types mainly through high SHAP score in monocytes and DCs. Barrier site IMIDs scored highest, with intriguing opposite directions in intestinal (high; CD and UC) and skin (low; PS ad PSA) diseases. We went on to validate *CYBA* overexpression in a PBMC validation cohort[49] and IBD patient biopsies using scRNA-seq[50]. The increased expression in monocytes as well as tissue-resident macrophages in IBD patients, suggests a role in the disease pathology, potentially through its oxidase function to produce superperoxide and ROS[68]. Importantly, UC mouse models treated with superoxide dismutase showed significantly attenuated UC disease burden in a dose-dependent manner and reduced lipid peroxidation in colonic tissue. Simultaneously, leukocyte rolling and adhesion in colonic venules of colitis rats were significantly reduced, contributing to strongly reduced inflammatory phenotypes[69].

Bringing reference atlases and data resources into the clinics is challenging without providing clear examples and strategies for their implementation. We generated a comprehensive landscape of inflammation in circulating immune cells from acute and chronic inflammatory diseases. Using advanced machine learning pipelines, we developed interpretable models for biomarker identification that can be further validated as a stand-alone or combinatory diagnostic test. On the other hand, we classified diseases based on generative models that learned the full inflammatory feature space across cell types and disease, laying the foundation for a universal diagnostic tool for inflammatory diseases.

Declarations

Ethics approval and consent to participate

Human blood processed in-house for this project was pre-selected and included within other ongoing studies. All the studies included were conducted in accordance with ethical guidelines and all patients provided written informed consent. Ethical committees and research project approvals for the different studies included in this manuscript are detailed in the following text.

ILCIC-D00 was approved by Hospital Universitari Vall d'Hebron Research Ethics Committee (PR(AG)144/201). **ILCIC-D01** received the IRB approval by the Parc de Salut Mar Ethics Committee (2016/7075/I). I**LCIC-D02** received the ethics approval by the Medisch-Etische Toetsingscommissie (METc) committe; for asthma patients (ARMS and ORIENT projects – NL53173.042.15 and NL69765.042.19 respectively), for COPD patients (SHERLOck project, NL57656.042.16), and finally, healthy controls (NORM project, NL26187.042.09). **ILCIC_D03** was approved by the Comité Ético de Investigación con Medicamentos del Hospital Universitario Vall d'Hebron (654/C/2019). **ILCIC-D06** was approved by the Comitè d'Ètica d'Investigació amb medicaments (CEim) del Hospital de la Santa Creu i Sant Pau (EC/21/373/6616 and EC/23/258/7364). **ILCIC-D12** was approved by the institutional review boards of the Commissie Medische Ethiek UZ KU Leuven/Onderzoek (S66460 and S62294).

Data and code availability

The complete raw data (FASTQ files) from in-house generated datasets as well as processed data will be available upon publication. The code to reproduce the full analysis is hosted in a private Github repository, and we will make it available upon publication.

Author information

LJG, JCN and HH conceived the project. JCN and HH supervised the project. LJG, DM, SAF, FC, MEB, MA, AS, RMB, HAT, RN, JN and AA performed the computational and statistical analysis. SR, DM, GC and YH generated datasets. TK, PvdV, FA, FP, MB, GS, ER, MvdB, AS, JMC, AFN, ED, JC, JT, JPG, AJ, VS, RE, ST, SV, MCN and SM provided patient samples. LJG, DM, SAF, FC, PV, ACV, JCN and HH interpreted the results. LJG, DM, SAF, FC, JCN and HH wrote the manuscript with input from all the authors. All authors read and approved the current version of the manuscript.

Acknowledgements

The authors would like to thank the helpful support received by the authors from publicly available data[70,71] used in the current study by providing processed data in the optimal format. Additionally, we appreciate the great effort put in creating the DISCO database[72] with multiple-tissue atlases using single-cell datasets, including human PBMCs. The authors would like to thank the CNAG Scientific IT Unit and the maintainers of the CNAG compute cluster for providing assistance with essential computing resources.

Funding

This project has received funding from the European Union's H2020 research and innovation program under grant agreement No. 848028 (DoCTIS; Decision On Optimal Combinatorial Therapies In Imids Using Systems Approaches). L.J.-G. has held an FPU PhD fellowship (FPU19/04886) from the Spanish Ministry of Universities. F.C. is founded by the Swiss National Science Foundation (SNSF) grant No CRSII5_205884/1. Y.H. is supported by a Junior Postdoctoral fellowship from the Research Foundation Flanders (FWO 12D5823N). S.T. is supported by a BOF-Fundamental Clinical Research mandate (FKO) from KULeuven and by the Belgian Foundation Against Cancer (FAF-C/2018/1301). V.S. is funded by Asociación Española Contra el Cáncer (AECC). M.C.N. acknowledges funding from GSK, the Netherlands Lung Foundation (project No. 4.1.18.226) and the European Union's H2020 Research and Innovation Program under grant agreement No. 874656 (discovAIR). This collaboration project is co-financed by the Ministry of Economic Affairs and Climate Policy by means of the PPP-allowance made available by the Top Sector Life Sciences & Health to stimulate public-private partnerships. A.S. is funded by PID2021-123918OB-I00 from MCIN/AEI/51 10.13039/501100011033 and co-funded by "FEDER: A way to make Europe".

Competing interests

H.H. is co-founder and shareholder of Omniscope, scientific advisory board member of Nanostring and MiRXES and consultant to Moderna and Singularity. J.C.N. is scientific consultant to Omniscope. V.S. has received research grants from AstraZeneca and honoraria from GSK unrelated to this study. M.v.d.B has received research grants (unrestricted) from AstraZeneca, Novartis, GlaxoSmithKline, Roche, Genentech, Chiesi and Sanofi. M.N. has been awarded with research grants (unrestricted) from AstraZeneca and GSK. A.S. is the recipient of research grants from Roche-Genentech, Abbvie, GSK, Scipher Medicine, Pfizer, Alimentiv, Inc, Boehringer Ingelheim and Agomab; receives consulting fees from Genentech, GSK, Pfizer, HotSpot Therapeutics, Alimentiv, Origo Biopharma, Deep Track Capital, Great Point Partners and Boxer Capital; and is on the advisory boards of BioMAdvanced Diagnostics, Goodgut and Orikine. A.A. is a computational biologist at IMIDomics, Inc. A.J. is the chief data scientist at IMIDomics, Inc. S.M. is the co-founder and CMO at IMIDomics, Inc.

References

1. Medzhitov, R. The spectrum of inflammatory responses. *Science* **374**, 1070–1075 (2021).
2. Newton, K., Dixit, V. M. & Kayagaki, N. Dying cells fan the flames of inflammation. *Science* **374**, 1076–1080 (2021).
3. Casanova, J.-L. & Abel, L. Mechanisms of viral inflammation and disease in humans. *Science* **374**, 1080–1086 (2021).
4. Agirman, G., Yu, K. B. & Hsiao, E. Y. Signaling inflammation across the gut-brain axis. *Science* **374**, 1087–1092 (2021).
5. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454**, 428–435 (2008).
6. Netea, M. G. *et al.* A guiding map for inflammation. *Nat. Immunol.* **18**, 826–831 (2017).
7. Roe, K. An inflammation classification system using cytokine parameters. *Scand. J. Immunol.* **93**, e12970 (2021).
8. Eltzschig, H. K. & Carmeliet, P. Hypoxia and Inflammation. *N. Engl. J. Med.* **364**, 656–665 (2011).
9. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. *FEBS J.* **285**, 2944–2971 (2018).
10. Soehnlein, O. & Libby, P. Targeting inflammation in atherosclerosis — from experimental insights to the clinic. *Nat. Rev. Drug Discov.* **20**, 589–610 (2021).
11. Psarras, A., Wittmann, M. & Vital, E. M. Emerging concepts of type I interferons in SLE pathogenesis and therapy. *Nat. Rev. Rheumatol.* **18**, 575–590 (2022).
12. Penkava, F. *et al.* Single-cell sequencing reveals clonal expansions of pro-inflammatory synovial CD8 T cells expressing tissue-homing receptors in psoriatic arthritis. *Nat. Commun.* **11**, 4767 (2020).
13. Manthiram, K., Zhou, Q., Aksentijevich, I. & Kastner, D. L. The monogenic autoinflammatory diseases define new pathways in human innate immunity and inflammation. *Nat. Immunol.* **18**, 832–842 (2017).
14. Zaiss, D. M. W., Pearce, E. J., Artis, D., McKenzie, A. N. J. & Klose, C. S. N. Cooperation of ILC2s and TH2 cells in the expulsion of intestinal helminth parasites. *Nat. Rev. Immunol.* 1–9 (2023) doi:10.1038/s41577-023-00942-1.
15. Cao, L. L. & Kagan, J. C. Targeting innate immune pathways for cancer immunotherapy. *Immunity* **56**, 2206–2217 (2023).
16. Dinarello, C. A., Simon, A. & van der Meer, J. W. M. Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. *Nat. Rev. Drug Discov.* **11**, 633–652 (2012).
17. Koenig, L. M. *et al.* Blocking inflammation on the way: Rationale for CXCR2 antagonists for the treatment of COVID-19. *J. Exp. Med.* **217**, e20201342 (2020).
18. Țiburcă, L. *et al.* The Treatment with Interleukin 17 Inhibitors and Immune-Mediated Inflammatory Diseases. *Curr. Issues Mol. Biol.* **44**, 1851–1866 (2022).
19. Taquet, M. *et al.* Acute blood biomarker profiles predict cognitive deficits 6 and 12 months after COVID-19 hospitalization. *Nat. Med.* **29**, 2498–2508 (2023).
20. Dann, E. *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* 1–11 (2023) doi:10.1038/s41588-023-01523-7.
21. Rauber, S. *et al.* Resolution of inflammation by interleukin-9-producing type 2 innate lymphoid cells. *Nat. Med.* **23**, 938–944 (2017).
22. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
23. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat.*

Methods **16**, 715–721 (2019).

24. Turner, M. D., Nedjai, B., Hurst, T. & Pennington, D. J. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochim. Biophys. Acta BBA - Mol. Cell Res.* **1843**, 2563–2582 (2014).
25. Pishesha, N., Harmand, T. J. & Ploegh, H. L. A guide to antigen processing and presentation. *Nat. Rev. Immunol.* **22**, 751–764 (2022).
26. Blum, J. S., Wearsch, P. A. & Cresswell, P. Pathways of Antigen Processing. *Annu. Rev. Immunol.* **31**, 443–473 (2013).
27. Bhat, M. Y. *et al.* Comprehensive network map of interferon gamma signaling. *J. Cell Commun. Signal.* **12**, 745–751 (2018).
28. Murayama, M. A., Shimizu, J., Miyabe, C., Yudo, K. & Miyabe, Y. Chemokines and chemokine receptors as promising targets in rheumatoid arthritis. *Front. Immunol.* **14**, (2023).
29. Lee, B.-W. & Moon, S.-J. Inflammatory Cytokines in Psoriatic Arthritis: Understanding Pathogenesis and Implications for Treatment. *Int. J. Mol. Sci.* **24**, 11662 (2023).
30. Kany, S., Vollrath, J. T. & Relja, B. Cytokines in Inflammatory Disease. *Int. J. Mol. Sci.* **20**, 6008 (2019).
31. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).
32. Reyes, M. *et al.* An immune-cell signature of bacterial sepsis. *Nat. Med.* **26**, 333–340 (2020).
33. Catalina, M. D., Bachali, P., Geraci, N. S., Grammer, A. C. & Lipsky, P. E. Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. *Commun. Biol.* **2**, 140 (2019).
34. Kalliolias, G. D. & Ivashkiv, L. B. Overview of the biology of type I interferons. *Arthritis Res. Ther.* **12**, S1 (2010).
35. Massoni-Badosa, R. *et al.* An Atlas of Cells in the Human Tonsil. 2022.06.24.497299 Preprint at https://doi.org/10.1101/2022.06.24.497299 (2022).
36. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* 1–12 (2023) doi:10.1038/s41587-023-01940-3.
37. Fernandez-Ruiz, R. & Niewold, T. B. Type I Interferons in Autoimmunity. *J. Invest. Dermatol.* **142**, 793–803 (2022).
38. Andreou, N.-P., Legaki, E. & Gazouli, M. Inflammatory bowel disease pathobiology: the role of the interferon signature. *Ann. Gastroenterol.* **33**, 125–133 (2020).
39. Dahlgren, M. W. *et al.* Type I Interferons Promote Germinal Centers Through B Cell Intrinsic Signaling and Dendritic Cell Dependent Th1 and Tfh Cell Lineages. *Front. Immunol.* **13**, (2022).
40. Golinski, M.-L. *et al.* CD11c+ B Cells Are Mainly Memory Cells, Precursors of Antibody Secreting Cells in Healthy Donors. *Front. Immunol.* **11**, (2020).
41. Mazzurana, L. *et al.* Crohn's Disease Is Associated With Activation of Circulating Innate Lymphoid Cells. *Inflamm. Bowel Dis.* **27**, 1128–1138 (2020).
42. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *The Lancet* **380**, 1590–1605 (2012).
43. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).
44. Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M. & Onishi, M. Multiobjective Tree-Structured Parzen Estimator. *J. Artif. Intell. Res.* **73**, 1209–1250 (2022).
45. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Preprint at https://doi.org/10.48550/arXiv.1907.10902 (2019).
46. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in*

Neural Information Processing Systems vol. 30 (Curran Associates, Inc., 2017).

47. Zhang, L., Yu, L., Li, J., Li, Z. & Zhao, X. Novel Compound Heterozygous CYBA Mutations Causing Neonatal-Onset Chronic Granulomatous Disease. *J. Clin. Immunol.* **43**, 1131–1133 (2023).
48. Denson, L. A. *et al.* Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* **154**, 2097–2110 (2018).
49. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).
50. Garrido-Trigo, A. *et al.* Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat. Commun.* **14**, 4506 (2023).
51. Gómez-Herranz, M., Taylor, J. & Sloan, R. D. IFITM proteins: Understanding their diverse roles in viral infection, cancer, and immunity. *J. Biol. Chem.* **299**, (2023).
52. Wen, L., Krauss-Etschmann, S., Petersen, F. & Yu, X. Autoantibodies in Chronic Obstructive Pulmonary Disease. *Front. Immunol.* **9**, (2018).
53. Hsu, A. T., Gottschalk, T. A., Tsantikos, E. & Hibbs, M. L. The Role of Innate Lymphoid Cells in Chronic Respiratory Diseases. *Front. Immunol.* **12**, (2021).
54. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
55. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
56. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
57. Edahiro, R. *et al.* Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nat. Genet.* **55**, 753–767 (2023).
58. DeMartino, J. *et al.* Single-cell transcriptomics reveals immune suppression and cell states predictive of patient outcomes in rhabdomyosarcoma. *Nat. Commun.* **14**, 3074 (2023).
59. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
60. Maina, E. K. *et al.* CD4+ T cell counts in initiation of antiretroviral therapy in HIV infected asymptomatic individuals; controversies and inconsistencies. *Immunol. Lett.* **168**, 279–284 (2015).
61. De Donno, C. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* 1–10 (2023) doi:10.1038/s41592-023-02035-2.
62. Youssef, A. *et al.* External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* 1–2 (2023) doi:10.1038/s41591-023-02540-z.
63. Sanyal, A. J. *et al.* Diagnostic performance of circulating biomarkers for non-alcoholic steatohepatitis. *Nat. Med.* **29**, 2656–2664 (2023).
64. van Steenbergen, H. W., Cope, A. P. & van der Helm-van Mil, A. H. M. Rheumatoid arthritis prevention in arthralgia: fantasy or reality? *Nat. Rev. Rheumatol.* 1–11 (2023) doi:10.1038/s41584-023-01035-y.
65. Feld, L., Glick, L. R. & Cifu, A. S. Diagnosis and Management of Crohn Disease. *JAMA* **321**, 1822–1823 (2019).
66. Clark, I. C. *et al.* HIV silencing and cell survival signatures in infected T cell reservoirs. *Nature* **614**, 318–325 (2023).
67. Zuroff, L. *et al.* Immune aging in multiple sclerosis is characterized by abnormal CD4 T cell activation and increased frequencies of cytotoxic CD4 T cells with advancing age. *EBioMedicine* **82**, 104179 (2022).
68. Jarmakiewicz-Czaja, S., Ferenc, K. & Filip, R. Antioxidants as Protection against Reactive

Oxidative Stress in Inflammatory Bowel Disease. *Metabolites* **13**, 573 (2023).

69. Seguí, J. *et al.* Superoxide dismutase ameliorates TNBS-induced colitis by reducing oxidative stress, adhesion molecule expression, and leukocyte recruitment into the inflamed intestine. *J. Leukoc. Biol.* **76**, 537–544 (2004).
70. Zhang, C. *et al.* Single-cell RNA sequencing reveals intrahepatic and peripheral immune characteristics related to disease phases in HBV-infected patients. *Gut* **72**, 153–167 (2023).
71. Palshikar, M. G. *et al.* Executable models of immune signaling pathways in HIV-associated atherosclerosis. *Npj Syst. Biol. Appl.* **8**, 1–15 (2022).
72. Li, M. *et al.* DISCO: a database of Deeply Integrated human Single-Cell Omics data. *Nucleic Acids Res.* **50**, D596–D602 (2022).
73. Cillo, A. R. *et al.* Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **52**, 183-199.e9 (2020).
74. Schafflick, D. *et al.* Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* **11**, 247 (2020).
75. Borcherding, N. *et al.* Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. *Commun. Biol.* **4**, 1–11 (2021).
76. Liu, Y. *et al.* Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nat. Commun.* **12**, 741 (2021).
77. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
78. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
79. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
80. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Annotated data. 2021.12.16.473007 Preprint at https://doi.org/10.1101/2021.12.16.473007 (2021).
81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
82. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
83. La Manno, G. *et al.* Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
84. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
85. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
86. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
87. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
88. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. Preprint at https://doi.org/10.12688/f1000research.8987.2 (2016).
89. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
90. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).

91. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
92. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
93. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. Preprint at https://doi.org/10.48550/arXiv.1802.03888 (2019).
94. Dong, W., Moses, C. & Li, K. Efficient k-nearest neighbor graph construction for generic similarity measures. in *Proceedings of the 20th international conference on World wide web* 577–586 (Association for Computing Machinery, 2011). doi:10.1145/1963405.1963487.
95. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification.

Methods

Atlas of Circulating Immune Cells

The Inflammation Landscape of Circulating Immune Cells (ILCIC) atlas has been conceived as a comprehensive resource to expand the current knowledge of physiological and pathological inflammation. With this aim, we have included data representing both acute and chronic inflammatory processes, as well as healthy donors. Further details about the included datasets are available (**Supplementary Table 1**).

Most of the single-cell RNA-sequencing (79%) has been generated in-house or shared by our collaborators from several research institutions. Samples were collected with written informed consent obtained from all participants and comply with the ethical guidelines for human samples. Specifically, we generated data from patients suffering Rheumatoid Arthritis (RA), Psoriatic Arthritis (PSA), Crohn's Disease (CD), Ulcerative Colitis (UC), Psoriasis (PS), Systemic Lupus Erythematosus (SLE) and healthy controls in collaboration with the Vall d'Hebron Research Institute within the DoCTIS consortia [https://doctis.eu/] (**ILCIC_D00**). Additionally, we processed and obtained data from healthy controls in collaboration with the Institut Hospital del Mar d'Investigacions Mèdiques (**ILCIC_D01**); Asthma, Chronic Obstructive Pulmonary Disease (COPD) and healthy control samples in collaboration with the University Medical Center Groningen (**ILCIC_D02**); Breast Cancer (BRCA) samples in collaboration with the Vall d'Hebron Institute of Oncology (**ILCIC_D03**); cirrhosis samples in collaboration with the Biomedical Research Institut Sant Pau (**ILCIC_D06**); and finally, samples of patients suffering Colorectal Cancer (CRC) in collaboration with the Katholieke Universiteit Leuven (**ILCIC_D12**).

Moreover, we also included public available datasets to complete our cohort. Therefore, raw count matrices and clinical metadata were obtained from the NCBI Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo/], the BioStudies Array Express [https://www.ebi.ac.uk/biostudies/arrayexpress] and Broad Institute DUOS [https://duos.broadinstitute.org/] resources. Specifically, we downloaded data for patients suffering sepsis[32] [SCP548] (**ILCIC_D04**), Head and Neck Squamous Cell Carcinoma (HNSCC)[73] [GSE139324] (**ILCIC_D05**), Hepatitis B Virus (HBV)[70] [GSE182159] (**ILCIC_D07**), Multiple Sclerosis (MS)[74] [GSE138266] (**ILCIC_D08**), clear cell Renal Cell Carcinoma (ccRCC)[75] [GSE121636, GSE121637] (**ILCIC_D09**), NasoPharyngeal Cancer (NPC)[76] [GSE162025] (**ILCIC_D10**), and Human Immunodeficiency Virus (HIV)[71] [GSE198339] (**ILCIC_D11**).

ILCIC barcodes. The ILCIC barcode was inspired by the TCGA project [https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/]. Each ILCIC barcode univocally identifies a cell, and it is composed by 5 alphanumeric identifiers representing project, data source, patient ID, library, and 10X Genomics cell barcode, respectively [e.g., ILCIC-D00-P006-L046-AAACCCAAGGTGAGAA].

Sample collection

Human blood samples were collected in EDTA tubes (BD Biosciences). Peripheral blood mononuclear cells (PBMCs) from the **ILCIC-D00, ILCIC-D02, ILCIC-D06** and **ILCIC-D12** datasets were isolated using Ficoll density gradient centrifugation (Lymphoprep™, Stem Cell Technologies; Ficoll-Plus, GE Healthcare Biosciences AB). PBMCs belonging to the **ILCIC-D01** and **ILCIC-D03** datasets were isolated using Vacutainer® CPT tube (BD Biosciences). Subsequently, all aliquots were centrifuged following the manufacturer's protocol. After centrifugation, PBMCs were washed and resuspended in freezing media. Aliquots were gradually frozen using a commercial freezing box (Mr. Frosty, Nalgene, Thermo Fisher Scientific) at -80 °C for 24 h before being transferred to liquid nitrogen for long-term storage.

Cell thawing and preprocessing

Cryopreserved PBMCs were thawed in a water bath at 37ºC and transferred to a 15 ml Falcon tube containing 10 ml of pre-warmed RPMI media supplemented with 10% FBS (Thermo Fisher Scientific). Samples were centrifuged at 350 x g for 8 min at RT, supernatant was removed and pellets resuspended with 1 ml of cold 1X PBS (Thermo Fisher Scientific) supplemented with 0.05% BSA (PN 130-091-376, Miltenyi Biotec). Samples were incubated during 10 min at RT with 0.1 mg/ml of DNAse I (PN LS002007, Worthington-Biochem) in order to eliminate ambient DNA and favor the resuspension of the pellet. Cells were filtered with a 40 µm strainer (PN 43-10040-70, Cell Strainer) to remove eventual clumps and washed by adding 10 ml of cold PBS+0.05% BSA. Samples were centrifuged at 350 x g for 8 min at 4ºC and resuspended in an adequate volume of PBS+0.05% BSA in order to reach the desired concentration. Cells concentration and viability were verified with a TC20™ Automated Cell Counter (Bio Rad) upon staining of the cells with Trypan Blue.

Sample multiplexing by genotyping

PBMCs samples were evenly mixed in pools of 8 donors per library following a multiplexing approach based on donor's genotype, as done by Kang *et al.*[77] for a more cost and time-efficient strategy. Importantly, in the case of **ILCIC-D00** libraries were designed to pool samples together from the same disease with different response to treatment (not relevant in this study) whereas in the case of the **ILCIC-D02 Asthma** cohort, 6 samples belonging to patients were pooled with 2 samples derived from non-smoking healthy control individuals. With this approach, we aimed to avoid technical artifacts that could mask subtle biological differences.

3' Cell Plex

PBMCs samples belonging to the **ILCIC-D01**, **ILCIC-D02 COPD**, **ILCIC-D06** cohort were multiplexed with 10X Genomics Cell Plex kit following the Cell Multiplexing Oligo Labeling for Single Cell RNA Sequencing Protocol (10x Genomics). Briefly, 0.2-1 million cells were centrifuged at 350x g at RT with a swinging-bucket rotor, resuspended in 100 µl of Cell Multiplexing Oligo (3' CellPlex Kit Set A PN-1000261, 10x Genomics) and incubated at RT for 5 mins. Cells were washed 3 times with cold 1X

PBS (Thermo Fisher Scientific) supplemented with 1% BSA (MACS Miltenyi), all centrifugations being performed at 350x g at 4C. Cells were finally resuspended in an appropriate volume of 1X PBS-1% BSA in order to obtain a final cell concentration of approximately 1600 cells/ul and counted using a TC20™ Automated Cell Counter (Bio-Rad Laboratories, S.A). An equal number of cells of each sample was pooled and filtered with a 40 µm strainer to remove eventual clumps, final cell concentration and viability of the pools were verified before loading onto the Chromium for cell partitioning.

Cell encapsulation and single cell RNA-sequencing library preparation

Multiplexed samples were loaded for a Target Cell Recovery between 20000 and 60000 cells (corresponding to 5000-7500 cells per sample within each plex). More specifically, samples belonging to **ILCIC-D00** and **ILCIC-D01** cohorts were encapsulated using standard throughput Chromium Next GEM Single Cell 3' Reagent Kit v3.1, while multiplexes samples belonging to **ILCIC-D02 Asthma and COPD,** and **ILCIC-D06** were encapsulated using the high throughput Chromium Next GEM Single Cell 3' HT Reagent Kit v3.1 in combination with the Chromium X instrument. On the other hand, samples of the **ILCIC-D03** and **ILCIC-D12** cohort were loaded in a standard assay with a target recovery of 6-8000 cells per sample using the Chromium Next GEM Single Cell 5' Reagent Kit v2 (10X Genomics, PN-1000263).

Libraries were prepared following manufacturer's instructions of protocols CG000315 or CG000390, for the standard assay without and with sample multiplexing, and protocols CG000416 and CG000419 for the high throughput assay without and with sample multiplexing. Protocol CG000331 was instead followed for the **ILCIC-D03** and **ILCIC-D12** cohort. Between 20-200 ng of cDNA were used for preparing libraries and final library size distribution and concentration were determined using a Bioanalyzer High Sensitivity chip (Agilent Technologies). Sequencing was carried out on a NovaSeq6000 system (Illumina) and NextSeq500 using the following sequencing conditions: 28 bp (Read 1) + 10 bp (i7 index) + 10 bp (i5 index) + 90 bp (Read 2), to obtain approximately 40.000 read pairs per cell for the Gene Expression (GEX) library and 2000-4000 read pairs per cell for the CellPlex library.

Data processing

To profile the cellular transcriptome, we processed the sequencing reads with the 10X Genomics Inc. software package CellRanger (v.6.1.1 for mostly all in-house generated datasets and v6.0.2 for CRC samples) [https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome] and mapped them against the human GRCh38 reference genome (GENCODE v32/Ensembl 98).

Genotype processing

Patient genotypes (VCF format) were simplified by removing Single Nucleotide Variants (SNVs) that were unannotated (chr 0), located in the sexual Y (chr 24), pseudo-autosomal XY (chr 25), or mitochondrial chromosomes (chr 26). As genotypes were obtained using the human hg19 reference genome, we converted their coordinates to the same reference genome used to mapped the sequencing reads (GRCh38), via the UCSC LiftOver tool [https://genome.ucsc.edu/cgi-bin/hgLiftOver].

LiftOver requires an input file in BED format. Thus, we used a python script [https://github.com/single-cell-genetics/cellsnp-lite/blob/master/scripts/liftOver/liftOver_vcf.py] to convert our VCF file accordingly.

Library demultiplexing

Multiplexed libraries from **ILCIC-D00** and **ILCIC-D02 Asthma** cohorts were demultiplexed with cellsnp-lite (v 1.2.2) in Mode 1a[78], which allows us to genotype single-cell GEX libraries by piling-up the expressed alleles based on a list of given SNPs. To do so, we used a list of 7.4 million common SNPs in the human population (MAF > 5%) published by the 1000 Genome Project consortium, and compiled by the authors [https://sourceforge.net/projects/cellsnp/files/SNPlist/]. Then, we performed the donor deconvolution with vireo (v 0.5.6)[79], which assigns the deconvoluted samples to its donor identity using known genotypes, while detecting doublets and unassigned cells. Finally, we discarded detected doublets and unassigned cells before moving on to the downstream processing steps. For CellPlex libraries, such as **ILCIC-D01**, **ILCIC-D02 COPD** and **ILCIC-D06**, we followed a joint deconvolution strategy combining CMO-hashing and genotype-based deconvolution; we generated pools of cells belonging to different samples based on the individual SNPs, and traced back to its donor of origin based on the CMO-hashing. When no genotype is available, the use of this dual approach minimizes the discarded cells.

Data analysis

All analyses presented in this manuscript were carried out using mainly Python. In particular, we structured our data in Anndata objects[80] compatible with Scanpy suite[81], which allowed us to apply single-cell data processing and visualization best practices. Details about the downstream analyses and the corresponding computational tools involved are detailed in the following method sections. Moreover, all findings presented in this manuscript are fully reproducible by running the code and notebooks uploaded in the project's Github repository. The package versions used in the analysis are reported at the end of each notebook.

Data standardization

Considering the diversity of the datasets included in the ILCIC atlas (project, source, chemistry, technology, Cell Ranger and genome reference version, and clinical metadata), a standardization step was needed.

Gene name harmonization. All datasets were mapped using human GRCh38 genome reference, but the annotation file version might differ, resulting in gene names with multiple aliases or deprecated symbols. Therefore, we compare all gene symbols with the HUGO Gene Nomenclature Committee (HGNC) database (latest version, July 2023) [https://www.genenames.org/], in order to convert them to the latest official HUGO symbol, merging possible duplicates.

Metadata harmonization. Patient metadata was unified across datasets, using common variable names and values for those present in multiple sources. For instance, '*M*', *'Male'*, and '*Hombre'* entries were replaced with '*male'.* Additionally, we created a new variable *'binned_age'* to group patients within a range of 10 years, considering that for the ILCIC-D01, ILCIC-D04, and ILCIC-D11 datasets the specific age information was not available.

Data quality control

We performed data Quality Control (QC) on the dataset count matrix by computing the main metrics (i.e., library size, library complexity, percentage of mitochondrial and ribosomal expression). Metric distributions were visualized grouping cells by library (10X Genomics) and by considering their chemistry (3' or 5' prime, and their version). Importantly, some datasets have been already processed by the original authors, therefore we assume that restrictive thresholds for cell/gene filtering have been already applied. Consequently, we removed low quality observation using permissive thresholds, while the robust cleaning process was performed during cell annotation tasks. In particular, we excluded the low quality libraries across datasets (<500 cells or <500 median genes recovered). Next, we removed cells with a very low number of UMIs (<400) and genes (<200), or with a high percentage of mitochondrial expression (>25%), as it is indicative of lysed cells. Then, we removed barcodes with a high library size (>50000 UMIs for 3' or >25000 UMIs for 5' chemistry) or with a high complexity (>6000 genes for 3' or >5000 genes for 5' chemistry), and also eliminated genes that were detected in less than 10 cells in the whole dataset. Lastly, we computed the cell cycle score using the gene list provided by the function *cc.genes.updated.2019()* from the Seurat library[82] (v 4.3.0.1), and defined the different cell cycle 'phases' (G1, G2M, and S).

Data processing for annotation

Annotation strategy. To identify all the immune cell types and states present in the human blood, we employed a recursive top-down approach inspired by previous work done by La Manno et al.[83] and Massoni-Badosa et al.[35]. Starting with more than 2M cells collected for the project, we divided the annotation into four steps. Briefly, in step 1 we grouped all cells into the primary compartments within our study. Subsequently, in step 2, each compartment was processed aiming to detect potential doublets, low quality cells and cells resembling platelets or erythrocytes (cells with high expression of hemoglobin genes). Additionally, we also placed back some clusters of cells into their main cell groups, when wrongly clustered due to similar profiles (e.g. T cells found into the NK cell group, or vice versa). In step 3, we identified the clusters resembling specific biological cell profiles (cell subtypes) obtaining a final number of 69 different subpopulations. Lastly, in the fourth step, cell identities were then hierarchically categorized in four levels of annotation, allowing for the automatic inference of broader levels of annotation. For each group identified in the previous steps, we applied the following tasks, namely: normalization, feature selection, integration, clustering and annotation. In the following, we will always refer to the parameters of step 1, while the specifics of the subsequent steps, along with the annotation labels and the marker genes used to define them, can be found **Supplementary Table 4**.

Data normalization. Following standard practices, filtered cells were normalized by total counts over all genes and multiplied by a scaling factor of 10^4 (*scanpy.pp.normalize_total(target_sum = 10^4)*). Then, the normalized count matrix *X* was log-transformed as $\log_e(X + 1)$ (*scanpy.pp.log1p()).*

Feature selection. Gene selection was performed by identifying the Highly Variable Genes (HVG). In order to reduce the influence of study's specific composition, and prevent biases in the gene selection task, we preferred genes that are highly variable in as many studies as possible. Therefore, similarly to Sikkema et al.[56], we first considered each study independently and computed the HVGs using the Seurat implementation[84] (*scanpy.pp.highly_variable_genes(min_disp=0.3, min_mean=0.01, max_mean=4)*). Then, we ranked genes based on the number of studies in which they are among the HV. Finally, for step 1, we determined the minimum number of studies required to compose a HVGs list of over 3000 genes. Applying this strategy, we selected a total of 3035 genes being highly variable in at least 5 studies.

Data integration. Our dataset includes single-cell data obtained via a number of different protocols, technologies, inflammatory status and a broad range of other clinical features (e.g., age, sex, etc). While this is a strength point of our atlas, such high levels of heterogeneity induced by technical confounding factors and unwanted biological variability resulted in challenging integration tasks before clustering and annotating cell populations. Therefore, we employed scVI[85], a Variational AutoEncoder (VAE) approach that proves to be one of the most effective integration methods in complex scenarios, particularly when the annotation information is missing[86]. scVI takes as input the raw count matrix to generate an integrated, lower-dimensional, embedded space where the cell states are preserved, and the batch effects are reduced. Moreover, scVI's lower dimensional space can be exploited to cluster and annotate cells based on either known or cluster specific marker genes. Regarding scVI's implementation, for step 1, we used 2 hidden layers for both encoder and decoder, a latent space of size 30 and the Negative Binomial (NB) as the gene likelihood distribution (*scvi.model.SCVI(adata, n_layers=2, n_latent=30, gene_likelihood="nb")*). We trained the model with early stopping on the VAE's ELBO loss function, using a patient of 50 epochs. For this step, scVI was trained for 413 epochs. As described in the following paragraphs, the integrated latent embedding generated by scVI was used for downstream analysis (clustering and visualization).

Cell clustering. In order to cluster cells into cell types with the Leiden algorithm[87], we first built the K-Nearest Neighbors (KNN) graph using scVI's latent embeddings and *k=20* as the number of neighbors (*scanpy.pp.neighbors(n_neighbors=k)*). We then applied the Leiden algorithm using a resolution of *r=0.05* (*scanpy.tl.leiden(resolution=r)*). The *k* and *r* used in every other step for every lineage can be found in **Supplementary Table 4.**

Cell annotation. Cell clusters were manually annotated by immunology experts by comparing the expression levels of canonical gene markers. Moreover, the final step of annotation (step 4) was performed using the clusters markers obtained performing a Differential Expression Analysis (DEA) among clusters (see **Supplementary Table 4)**. First, we ranked genes to characterize each cluster (*scanpy.tl.rank_genes_groups()*), by considering normalized RNA counts with the Wilcoxon sum rank

test. Then, we selected those genes with Log2 Fold Change (Log2FC) > 0.2, and with a p-value < 0.05. Notice that p-values were corrected by applying the False Discovery Rate (FDR) approach.

External annotation validation. As an additional validation, we compared our independent annotations with the ones available in the public datasets. To quantify the overlap of cells among groups, we computed the pairwise Jaccard Index between each cell identifier grouped by the external annotation and each cell identifier grouped using our internal annotations (level 2) (**Supplementary Table 2**).

Feature selection post annotation

Gene selection. A gene selection step is necessary to improve the quality of downstream analysis to characterize the inflammation landscape. In order to remove dataset specific genes and reduce the batch effect, we first performed data normalization (as described above) and then removed all the genes that are not expressed (raw count > 0) in at least 1 cell in each study. This step retained a total of 12446 genes. Then, we identified three sets of genes: (i) the HVGs, (ii) the Differentially Expressed Genes (DEGs) between healthy and each inflammatory status, and (iii) a manually curated immune-specific gene list. Importantly, cells belonging to Hemoglobin and Platelet populations were excluded from all the downstream analyses, except for label transfer performed in the patient classifier (as explained below).

Highly variable genes (HVGs). Similarly to the feature selection approach described in the annotation step, we selected a total of 2330 HVGs, by using a threshold of at least 2000 genes. In practice, we first ranked the genes based on the number of studies in which they are concurrently highly variable (*scanpy.pp.highly_variable_genes(min_disp=0.3, min_mean=0.01, max_mean=4, batch_key='ILCIC_library'))*, and then chose a minimum number of studies of 3.

Differentially Expressed Genes (DEGs) between healthy and each disease. As suggested by standard practice, the suggested approach to extracted DEGs is to compute them after grouping single-cell gene expression profiles into pseudobulks. Therefore, we first combined the expression profiles to produce pseudobulks for every patient and cell-type (annotation level 1), removing groups with no more than 20 cells, using the Python implementation of decoupleR[31] (*decoupler.get_pseudobulk(min_cells=20,sample_col='ILCIC_patient',groups_col='annotation_depth02', layer='counts', mode='sum')*). Then, we applied the edgeR's[88] quasi-likelihood functions to search for DEGs between healthy patients and each other's inflammatory conditions, by considering one cell-type at a time. Since not all the cell-types were detected in each patient, we didn't perform the pairwise comparison if one disease had less than 3 pseudobulks. More in detail, for each pairwise comparison we first removed genes with a low expression value (*filterByExpr(y, group = disease)*). Second, we normalized by library size the aggregated raw counts (*calcNormFactors(y, logratioTrim = 0.3)*). Third, we corrected for the main confounding factor, that is, the sequencing chemistry (i.e., 5', 3' v.2, and 3' v.3), and also for patient gender, considering an additive model. We defined the design of our comparison using the following patsy-style [https://patsy.readthedocs.io/en/latest/formulas.html] formula: '*~0 + C(disease) + C(mainChemistry) + C(sex)*'. Fourth, we estimated a Negative Binomial

(NB) dispersion for each gene using (*estimateDisp()*) which we feed into a gene-wise NB Generalized Linear Model (*glmQLFit(robust=TRUE)*) to test for differentially expressed genes with a quasi-likelihood F-test (*glmQLFTest()*). Lastly, results obtained from each comparison, were merged together and the F-test p-values were corrected using the Benjamini-Hochberg False Discovery Rate procedure implemented in R (*p.adjust(method = 'BH')*). Given the corrected p-values and the $\log_2$-fold change ($\log_2$FC) we selected 2296 DEGs with p-value < 0.01 and absolute $\log_2$FC > 2.

Manually curated immune-specific genes. To be able to track the full spectrum of an inflammatory process, including immune activation and progression, we ensure that the genes present in our human immune-specific genes curated list (491 genes, only 451 present in our dataset; **Supplementary Table 3**)[24–30], defined by an immunologist expert and based on the literature, were also included. Such genes are grouped in 14 inflammation related functions, as reported in the above Table.

Aggregation of gene sets. We generate the relevant gene set by doing the union of HVGs, DEGs, and the manually curated list. The final number of unique genes is 3986.

Datasets integration and gene expression correction via scGen

Atlas-level analysis requires a careful preprocessing of the gene expression profiles to deal with the heterogeneity of the studies, the batch effect and the missing or noisy observations[89]. scGEN[23] is one of the existing methods that is able to tackle these challenges, and was also proven to be effective on atlas-level benchmarks against other methods that provide corrected expression matrices[86].

scGen integration. scGen is defined by two main components: a Variational AutoEncoder (VAE) and a latent space arithmetic method. The VAE estimates a posterior distribution of latent variables through the encoder, from which we can reconstruct the expression matrices via the decoder (*scGen_model.batch_removal()*). Similarly to commonly employed VAEs, scGen approximates the posterior through a variational distribution, modeled by the encoder and defined as a multivariate Gaussian. When the scGen's VAE has been fitted, latent space arithmetic is employed to correct for the batch effect induced by the technology (*ILCIC-technology*). Within each cell type, scGen first selects the mean μ_{max} of the most populated batch, and then corrects each batch with mean μ_0 by adding $\delta = \mu_{max} - \mu_0$ to each cell's embedding. Importantly, the cell type has to be inferred when not known, as done in the patient classifier described in the next section. The final corrected count matrix will correspond to the generated count matrix from the arithmetic-corrected embeddings. Following scGen's tutorials, we will refer as corrected embeddings to the ones obtained given the corrected expression matrix as input. We trained scGen with 200 as latent dimension and early stopping with a patient of 10 epochs (*scgen.SCGEN(n_latent*=200, *early_stopping*=True*).train(early_stopping=True,* early_stopping_patience=10*))*. Lastly, the batch effect taken into consideration was the technology "*ILCIC-technology*" and the employed annotation level was the level 2.

Comparison of gene expression profiles

Compositional cell-type analysis. To estimate the changes in the proportions of cell populations across diseases, we applied the scCODA Python package[90], a Bayesian modeling tool that takes into

account the compositional nature of the data to reduce the risk of false discoveries. scCODA allows us to infer changes between conditions while considering other covariates, corresponding to the disease status in our setting. scCODA searches for changes between a reference cell type, assumed to be constant among different conditions, and the other cell types. In order to run our analysis in an unsupervised way, we let scCODA automatically define such a reference. scCODA takes as input the count of cells belonging to each cell type in each patient and returns the list of cell types proportion changes and the corresponding corrected p-values (through the False Discovery Rate procedure; FDR). A *patsy-style* formula was used to build the covariate matrix, specified with 'healthy' as baseline (*C(disease, Treatment('healthy'))*), since we are interested in detecting changes between a normal and diseased status. Finally, the model was fitted by running *sccoda.util.comp_ana.CompositionalAnalysis(formula, reference_cell_type = "automatic"*) and we consider as relevant only the changes with a corrected p-value lower than 0.05 and a log2-FoldChange higher than 0.2.

Inflammation-related signature scores. To compare immune-relevant activation profiles across diseases and cell types, we applied an enrichment signature scoring procedure considering inflammation-related functions from the manually curated immune-specific genes previously introduced. In particular, we applied the Multivariate Linear Model (MLM) approach using the Python implementation of DecoupleR[31], which has been proven to return more reliable results compared to other scoring methods. Starting from the scGen correct gene expression matrix, we reduced the impact of highly expressed genes, standardizing their expression values by removing the mean and scaling them to unit variance. Moreover, we considered only inflammation functions that included at least 5 genes, resulting in 12 signatures. The last filtering step involves excluding cells belonging to HBcells, Platelets and HSC cell types (level 1), as they contain very few cells and are not present across all diseases. We fitted decoupleR's MLM (*decoupler.run_mlm()*) by considering the pre-processed input as response variables and the signature gene sets with unitary weights as covariates. The output of the model is a t-student statistic for each cell and each inflammation function, which is used as a proxy of its activation score. Thus, positive values are associated with more active functions in a given cell, compared to the other cell functions, while negative values refer to functions less active. Finally, we compared the activation score changes between *'healthy'* cells and each disease in three resolution levels: i) ungrouped cells, ii) cells grouped by annotation level 1, or iii) annotation level 2. In the latter resolution level, we consider only groups with at least 20 cells. More in detail, we first computed the average activation score in each given cell group. Then, we considered the 'healthy' score as reference values (h) and computed the Mean Relative Percentage change between with diseased ones (d_i) as following: $((d_i - h) / |h|) * 100$.

Gene factor inference. In order to validate our curated list of immune-related genes, we employed Spectra[36], a tool able to identify a minimal set of genes related to specific functions in the data (i.e., factors). Given our dataset, we expect Spectra to select gene modules that are related to inflammatory functions. More in detail, we started our analysis with the raw count expression values of our selection of ~4k genes and the curated list of immune-related functions with the corresponding gene sets. We further filtered the input count matrix by removing 103 genes related to T and B cell receptors which are

strongly differentially expressed among subclones. Therefore, we reduced the impact of such heterogeneity in our analysis. Moreover, we kept only immuno-related functions that included at least 5 genes, following Spectra's tutorial. Additionally, due to the high resource requirements of Spectra, we computed the pseudobulks of each cell type in each patient in order to reduce the sample size. In particular, we applied the decoupleR Python library (*dc.get_pseudobulk(spectra_adata,sample_col='ILCIC_patient',groups_col='annotation_depth04_21Sep23', layer='counts',min_cells=20, min_counts=0)),* and obtained the pseudobulks on level 2 annotations. Next, we normalized obtained profiles by applying the same procedure explained before (see **Data normalization** section). Then, the Spectra model was fitted with default hyperparameters except for lambda, which was set to 0.05 to reduce the importance of the curated gene sets, given more flexibility to Spectra's inference (*Spectra.est_spectra(lam=0.05)).* Among the factors returned by Spectra, we focused on the ones related to our *IFN_response* function provided as input. In particular, we chose the IFN-related factors identified in each cell type, by selecting the ones which obtained an overlap coefficient higher than 0.4 between the genes in the original *IFN_response* set and the top 50 genes ranked by their score in each factor. Moreover, Spectra's output also includes a matrix with an activity score for each factor in each sample (i.e., pseudobulk). Starting from such a matrix, we first averaged the activity scores by grouping the samples by disease. Then, we scaled each factor by dividing each score by the max value among the diseases. Finally, we run a Principal Component Analysis implemented in the Scanpy library (*scanpy.pp.pca()*).

Immune gene importance evaluation

In this section, we will introduce our pipeline to obtain a gene importance metric by interpreting cell-type-specific multi-class classifiers for disease prediction. In particular, classifiers are based on Gradient Boosted Decision Trees (GBDTs) implemented in the CatBoost library[91], which has been shown to provide the best performance among existing GBDTs[43]. Lastly, interpretability was performed using SHapley Additive exPlanation (SHAP) values[46], a powerful approach taking into account also the interactions between the genes to assign the importance score.

CatBoost Fitting. The first step of the pipeline consists in extracting the corrected expression from scGen as described in the previous section. Then we standardized the expression by removing the mean and scaling to unit variance (*sklearn.preprocessing.StandardScaler()*); this step was necessary in order to have equal features scale for L2 regularization employed by CatBoost GBDTs. Due to the large size of our datasets, we did not consider a full cross-validation but a simpler train-validation-test splitting, where the test set accounts for 20% of the data and the validation set for 10% of the training set. The validation set was used for CatBoost (*CatBoostClassifier(eval_metric='TotalF1:average=Macro',max_iterations=2000,bootstrap_type="Bayesian")*) early stopping (with a patient of 50 iterations), while performance was evaluated on the test set. Hyperparameters were tuned through the Optuna library[45], which employs a Tree-structured Parzen Estimator (TPE) sampling algorithm[44] to navigate the hyperparameters space. We considered the following hyperparameters: *l2_leaf_reg*={2, 3, …12}, *depth*={4, 5, …10}, *random_strength*={0, 1, …, 10}, *colsample_bylevel*={0.01, 0.1}, *baggin_temperature*={0, …, 10}. We allowed Optuna to test 100

different configurations, while employing pruning techniques (*optuna.pruners.MedianPruner(n_warmup_steps=100, n_startup_trials=5)*) to prune unpromising runs and reduce computational time. In order to evaluate the performance of a configuration, we employed F1-score with macro averaging. The above fitting procedure was applied independently on each cell type belonging to level 1 (without HSC, pDC, Platelet, and HBcell).

SHAP interpretability. While Machine Learning approaches such as decision trees or Naive Bayes are inherently interpretable, GBDTs require post-hoc interpretability tools in order to infer explanations. In particular, explanations can be global, if they convey the behavior of the model on the whole data distribution, or local, if it regards a single input sample. Post-hoc interpretability tools for GBDTs include techniques that are either global or local, such as permutation-based feature importance[92,93]. Recently, a new approach based on game-theory, namely SHapley Additive exPlanation (SHAP) values[46], has been shown to provide explanations that are locally consistent and can also be globally aggregated. SHAP values explain the output of the classifier, in our case the predicted disease, as a sum of contributions of each feature, that is, genes. Such contributions correspond to the change in the expected model prediction when conditioning on the considered feature. Consequently, the importance of a gene is related with the expected impact that a change in its expression has on the confidence of the classifier towards predicting a specific disease. In other words, a gene gets a low importance value if changing its expression leads to a small prediction change (and vice versa). Crucially, the order in which genes are considered is meaningful due to the interactions between them, and the optimal SHAP value for each gene is the average effect given all possible orderings. Computing the optimal SHAP values requires evaluating an exponential number of gene orderings. However, for tree-based methods, SHAP values can be computed in polynomial time[93]. Therefore, we computed the SHAP values through the implementation provided by the CatBoost library (*CatBoostClassifier.get_feature_importance(type="ShapValues")*). At the end of the pipeline and given a specific cell type *ct*, we end up with a SHAP value for every gene in every cell, and for each disease: a matrix of real values $shap^{ct}(c, g, d)$, where *c, g,* and *d* identify the cell, gene, and disease, respectively. The average contribution of a gene *g* for a disease *d* can be computed as $Dshap^{ct}(g,\ d)\ =\ mean_{c\,\in C}|shap^{ct}(c, g, d)|$, where *C* is the set of cells. While the global importance of a gene *g* is $GCshap^{ct}(g)\ =\ sum_{d\,\in D}Dshap^{ct}(g, d)$, where *D* is the set of diseases. By repeating a similar analysis across classifiers (and cell types), we can get the global importance of a gene in a specific disease *d*. More in detail, we can compute the importance of gene *g* in a cell-type *ct* as $Cshap^{d}(g,\ ct)\ =\ mean_{c\in C}|shap^{ct}(c, g, d)|$ such that *C* is the set of cells, and the global importance across cell types as $GDshap^{d}(g)\ =\ sum_{ct\,\in CT}Cshap(g, ct)$, where *CT* is the set of all cell types. Note that while $Dshap^{ct}(g,\ d)$ and $GCshap^{ct}(g)$ are a standard interpretation of a classifier with SHAP, $Cshap^{d}(g,\ ct)$ and $GDshap^{d}(g)$ integrate the importance scores *across* classifiers.

Patient classifier

In this section, we define our pipeline to predict a patient's disease status from their single-cell gene expression profiles. This pipeline is expected to work with an annotated training dataset with known cell type (i.e. our atlas, and a test dataset without annotations, new patient data). We validated our pipeline by simulating a reference atlas and an unseen dataset by splitting our data into a train (80% patient) and test (20% patients) split. Then, we removed the available annotations from the latter. The described steps are then applied 5 times through cross-validation.

Extracting scGen embeddings. Starting from the training dataset, we first train scGen to obtain a latent embedding as described in the previous section. This way, we reduce the dimensionality of each cell to scGen's latent size (200). Importantly, a label transfer step, described in the next paragraph, is required to apply scGen's latent space arithmetic on the test set, since the cell type annotation is required to correct the embedding.

Label transfer. In order to apply scGen batch correction on the test set, we need to also infer the cell types of those cells. This step was performed through label transfer by nearest neighbors, following a similar approach employed in Human Lung Cell Atlas[56] and introduced in[55]. The idea is to employ (approximate) nearest neighbors through the PyNNDescent[94] (*pynndescent.NNDescent().prepare()*), and infer the most probable cell type in the 10 nearest neighbors (*pynndescent.NNDescent().query()*) from the already annotated cells in the training set. To account for the shape of the distribution of the neighbors, a Gaussian kernel was applied instead of using the Euclidean distance. The most probable nearest neighbor cell type is then assigned to annotate new cells. Since both training and test set belong to our annotated dataset, we evaluated the label transfer goodness by computing the balanced accuracy score, comparing the predicted annotation in each test set with the previous assigned ones.

Embedding correction. Having the annotations for both the training and test sets, we can now correct both of them by applying scGen's latent space arithmetic. In particular, we first extracted □s from the training set batch correction (see **scGen integration** section) and add them to the test cell embedding, in order to have cells in a common, corrected latent space.

Pseudobulk generation. To account for the still large and variable-sized feature spaces, we applied pseudobulk through decoupleR[31] among patients and grouping by cell type (annotation level 1), excluding HSC, Platelet, and HBcell. This step assigns to each patient a pseudobulk to each defined cell type, given that at least one cell for the cell type was detected. Then, a Linear Support Vector Classifier[95] (*sklearn.svm.LinearSVC(max_iter=10000, dual=True)*) in a one-vs-rest multi-class strategy, such that each learned decision boundary separates each class from all the others. Moreover, to deal with the case of having a missing cell type in a given patient, we proposed to apply the LinearSVC classifier independently to each cell type, and then aggregate the prediction with a majority voting approach. Thus, the aggregated prediction corresponds to the most frequent prediction among each cell type's classifier. In case the most frequent disease is not unique, the patient is labeled as '*undetermined*'. Lastly, we excluded ccRCC disease from the analysis, due to the low number of patients included.