Journal, 2024, pp. 1–13

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

singIST: an integrative method for comparative single-cell transcriptomics between disease models and humans

Aitor Moruno-Cuenca,[1,2,3,4,*] Sergio Picart-Armada,[1] Alexandre Perera-Lluna[2,3,4] and Francesc Fernández-Albert[1]

[1]Data Science, R&D Center, Almirall SA, Sant Feliu de Llobregat, Spain, [2]B2SLab, Institut de Recerca i Innovació en Salut (IRIS), Universitat Politècnica de Catalunya, Barcelona, Spain, [3]Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain and [4]Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain

[*]Corresponding author. aitor.morunocuenca@almirall.com
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Disease models serve as fundamental tools in drug discovery and early-stage drug development. However, these models are not a perfect reflection of human disease, and selecting a suitable model can be challenging. Existing computational approaches for molecular validation of pathophysiological resemblance to human conditions at single-cell resolution remain limited. Although quantitative computational methods exist to inform this selection, they are very limited at the single-cell resolution, which can be critical for model selection. Quantifying the resemblance of disease models to the human condition with single-cell technologies in an explainable, integrative, and generalizable manner remains a significant challenge.

Results: We present singIST, a computational method for comparative single-cell transcriptomics analysis between disease models and human conditions. singIST provides explainable quantitative measures on disease model similarity to human condition at both pathway and cell type levels, highlighting the importance of each gene in the latter. These measures account for orthology, cell type presence in the disease model, cell type and gene importance in human condition, and gene changes in the disease model measured as fold change. This is achieved within a unifying framework that controls for the intrinsic complexities of single-cell data. We tested our method using three well-characterized murine models of moderate-to-severe Atopic Dermatitis, demonstrating its ability to recapitulate established biological knowledge while generating novel hypothesis through pathway-level analysis.

Availability and implementation: Source code at `https://github.com/amoruno/singIST-reproducibility`

Key words: Computational biology, Machine learning, Translational medicine, Single Cell, Comparative transcriptomics

1. Introduction

Disease models are biological experimental systems to study human disease. These models are designed to mimic the pathophysiology, progression, and response to treatments observed in human conditions. These models serve as the backbone to drug discovery and early drug development activities; drug target validation and characterization (Emmerich et al., 2021); compound screening (Elitt et al., 2018; Wei et al., 2021); preclinical studies to identify a lead candidate from several targets, select optimal formulation, posology and route of administration (Shegokar, 2020); guide early phase clinical trial design (Steinmetz and Spack, 2009; Loewa et al., 2023). However, the validation of molecular physiology, etiology and pathogenesis of disease models to that of human condition remain a challenge, which contribute to high rates of drug development attrition (Storey et al., 2022).

Recently, there have been methodological advancements in bioinformatics to quantitatively assess the validity of disease models in mimicking a human condition, through bulk transcriptomics. Found In Translation (FIT) (Normand et al., 2018) is a statistical methodology, relying on regularized linear regression models, that leverages bulk transcriptomics data to extrapolate murine disease models' gene expression to expression changes that would be equivalent in the human condition, by using disease models' Fold Changes (FC). Another approach is In Silico Treatment (IST) (Picart-Armada et al., 2024), a computational method that assesses translation

of disease-related bulk gene expression patterns between animal models and humans, by also simulating observed disease models' FC onto humans, providing an interpretable measure of their transcriptomics similarity. Nonetheless, evaluating disease models using bulk transcriptomics methods may lack the necessary granularity to underpin changes in specific cell populations involved in the pathological manifestations of the human condition. This is particularly true for Immune-mediated inflammatory diseases (IMIDs), whose pathogenesis is primarily driven by lymphoid cells (McInnes and Gravallese, 2021; Pisetsky, 2023). FIT nor IST provide a trivial approach to accommodate for single-cell data. Current methodologies for comparative analysis of single-cell transcriptomic changes in disease models' to that of human condition are scarce. A recurrent approach is to perform an Overlapping Differentially Expressed Genes (ODEGs) analysis between disease models and human condition (Kim et al., 2019; Li et al., 2023; Ali et al., 2024), yet ODEGs have been proven to be suboptimal as it treats every gene direction and magnitude as equal posing the need for more sophisticated approaches (Lawhorn et al., 2018). Another strategy is performing a dimensionality reduction technique (CCA, NNMF, tSNE) on disease models' and human scRNA-seq data and compare the obtained latent factors (Gao et al., 2021; Karmele et al., 2023; Franzén et al., 2024), which poses difficulty in interpreting and quantifying the similarity between both.

To address the challenges in single-cell transcriptomics analysis, we introduce singIST, a flexible computational method built on the foundation of IST. singIST facilitates comparative analysis between disease models and human conditions by accounting for orthology, cell type agreement, adaptive sparsity, and the importance of genes and cell types. It provides interpretable measures of transcriptomic similarity at different levels of granularity. We demonstrate its potential by assessing well-characterized murine models of Atopic Dermatitis (AD), a skin IMID, focusing on deregulated biological pathways.

2. Materials and methods

2.1 Materials

2.1.1 Human data

Human moderate to severe AD scRNA-seq data correspond to patients diagnosed with chronic AD in early childhood, obtained from all Healthy Control (HC) and AD skin suction blisters samples analyzed in Bangert et al. (2021). In total, 4 HC and 5 AD skin suction blisters were used. Metadata and GEO identifiers can be found in Supplementary Material S4. Cell type populations modelled are those identified by Bangert et al. (2021): T-cells, Melanocytes, Dendritic Cells, Langerhans Cells and Keratinocytes. Raw counts were pseudobulked and posteriorely log normalized using Seurat v5.0.1. Human scRNA-seq were obtained by 10x Genomics sequencing platform.

2.1.2 Disease model data

We evaluate scRNA-seq of three epicutaneous sensitized murine models that cause and AD like eczema phenotype; Oxazolone (OXA) and Imiquimod 5% cream (IMQ), obtained from Liu et al. (2020); Ovalbumine (OVA), obtained from Leyva-Castillo et al. (2022), the latter two are also stablished murine models in Psoriasis and Asthma, respectively. Metadata and GEO identifiers can be found in Supplementary Material S4. All disease models, and their respective controls, have 3 replicates of ear skin biopsies. To allow for comparison with human data, the existing cell type annotations of disease models were mapped to match those observed in the human dataset, the mapping of this relation is shown in Supplementary Material S2. Likewise to human samples, disease model samples were pseudobulked and log normalized thereafter. All disease models scRNA-seq were obtained by 10x Genomics sequencing platform.

2.1.3 Pathway data

Pathways under analysis were selected from Brunner et al. (2017), all enriched pathways in human moderate to severe AD compared to HC in serum. Gene sets were retrieved from MsigDB version 7.5 (Liberzon et al., 2015), pathway database source encompasses: KEGG (Kanehisa et al., 2023), REACTOME (Gillespie et al., 2022), BioCarta (Nishimura, 2001), PID (Schaefer et al., 2008). Only curated pathways from MsigDB were used (C2), those archived by MsigDB were excluded. In total, 22 gene sets satisfying the former criteria were considered.

2.2 Methods

2.2.1 singIST method

We start by defining the three inputs of singIST: superpathways, human scRNA-seq, and disease model scRNA-seq Fold Changes (FC). First, we define the concept of a superpathway $\mathcal{P}^p$ as a set containing cell types and genes. We name $C^1, \ldots, C^b, \ldots, C^B$ as the cell types of interest, previously identified and annotated. For each superpathway $\mathcal{P}^p$, there is a gene set $\mathcal{G}_p$ extracted from a pathway of interest p, from which gene subsets are derived for the cell types $\mathcal{G}_p^1, \ldots, \mathcal{G}_p^b, \ldots, \mathcal{G}_p^B$ where $\mathcal{G}_p^b \subseteq \mathcal{G}_p \ \forall b$. Each superpathway is formally defined as $\mathcal{P}^p = \bigcup_{b=1}^{B} \mathcal{G}_p^b$, with the complete set of superpathways $\mathcal{P} = \{\mathcal{P}^1, \ldots, \mathcal{P}^p, \ldots, \mathcal{P}^P\}$ representing all pathways under evaluation. singIST method runs independently for each of the superpathways; hence, without loss of generality, from now on we fix a superpathway $\mathcal{P}^p$.

Second, we structure the human scRNA-seq data according to the superpathway. Let $\underset{\sim}{C} = [\boldsymbol{C}^1, \ldots, \boldsymbol{C}^b, \ldots, \boldsymbol{C}^B]$ be the block of matrices containing the pseudobulk log-normalized expression for each cell type. Each matrix is defined element-wise $\boldsymbol{C}^b = \{x_{ig}^b\}_{\substack{1 \le i \le n \\ g \in \mathcal{G}_p^b}}$, where x_{ig}^b is the pseudobulk gene expression of human sample i for gene g in cell type b. We assume that human samples belong to different experimental groups $k = 1, \ldots, K$, from which we define the *target class* $k = k_1$ as the human experimental group that the disease model is intended to mimick (i.e disease, treated, etc.), and the *base class* $k = k_0$ as the human experimental group that should differentiate from the *target class* (i.e healthy control, untreated, etc.). Let $\boldsymbol{Y}$ be the response matrix denoting human sample class, defined element-wise $\boldsymbol{Y} = \{Y_{ik}\}_{\substack{1 \le i \le n \\ 1 \le k \le K}}$, where elements Y_{ik} satisfy (1):

$$Y_{ik} \in \{0, 1\}, \forall (i, k) \in \{1, \ldots, n\} \times \{1, \ldots, K\}$$

$$\sum_{k=1}^{K} Y_{ik} = 1, \forall i \in \{1, \ldots, n\} \tag{1}$$

Third, we assume there are $l = 1, \ldots, L$ disease models to be assesed against human scRNA-seq data for each superpathway. Since singIST runs independently for each disease model, without loss of generality we fix a disease model l. We structure

	Notation and Symbol	Description
Symbol	$\widehat{e}$	Estimation of element e
	e'	Element e for human singIST treated samples
	$\underset{\sim}{e}$	Block of vectors or matrices e
	e^i	Element i of $\underset{\sim}{e}$, or variable containing information thereof
Notation	C^b	Cell type of interest b
	$\mathcal{G}_p$; $\widetilde{\mathcal{G}}_p$	Gene set of pathway p; Equivalent of $\mathcal{G}_p$ for disease model organism gene symbol
	$\mathcal{G}_p^b$; $\widetilde{\mathcal{G}}_p^b$	Gene subset of pathway p for cell type b, $\mathcal{G}_p^b \subseteq \mathcal{G}_p$; Equivalent of $\mathcal{G}_p^b$ for disease model
	$\mathcal{P}^p$	Superpathway $\mathcal{P}^p = \bigcup_{b=1}^{B} \mathcal{G}_p^b$
	x_{ig}^b	Human pseudobulk of gene $g \in \mathcal{G}_p^b$, sample i and cell type b
	$\boldsymbol{C}^b$	Matrix with human pseudobulk for cell type b, $\boldsymbol{C}^b = \{x_{ig}^b\}_{\substack{1\leq i\leq n \\ g\in\mathcal{G}_p^b}}$
	Y_{ik}	A binary variable that is 1 if human sample i is in class k, and 0 otherwise
	$\boldsymbol{Y}$	Matrix with human class $\boldsymbol{Y} = \{Y_{ik}\}_{\substack{1\leq i\leq n \\ 1\leq k\leq K}}$
	$k = k_1$; $k = k_0$	*Target class*: human experimental group that the disease model should mimick; *Base class*: human experimental group with assumed different scRNA-seq profile from the *target class*
	$r_{\tilde{g}}^b$	Disease model FC between k_1 and k_0 for disease model gene $\tilde{g} \in \tilde{\mathcal{G}}_p^b$
	$\boldsymbol{R}^b$	Vector with disease model FC for cell type b, $\boldsymbol{R}^b = \{r_{\tilde{g}}^b\}_{\tilde{g}\in\tilde{\mathcal{G}}_p^b}$
	y_{ik}	Superpathway's score of asmbPLS-DA for sample i and class k
	Ω_k	Difference between the median values of y_{ik} for class k and k_0
	γ_{ik}^b	Cell type b contribution to y_{ik}
	Γ_k^b	For cell type b, difference between the median values of γ_{ik}^b for class k and k_0
	δ_{igk}^b	For cell type b, gene $g \in \mathcal{G}_p^b$ contribution to γ_{ik}^b
	Δ_{gk}^b	For cell type b, difference between $\delta_{igk}^b{}'$ and $\delta_{igk_0}^b$, which is constant for all samples i
	Ωf_k	Ω_k' as a fraction of Ω_k
	Γf_k^b	$\Gamma_k^b{}'$ as a fraction of Γ_k^b
	Δf_{gk}^b	Δ_{gk}^b as a fraction of Γ_k^b

Table 1. Summary table of main symbols and notations defined in singIST method.

the disease model scRNA-seq FC as blocks of vectors. Let $\underset{\sim}{R} = [\boldsymbol{R}^1, \ldots, \boldsymbol{R}^b, \ldots, \boldsymbol{R}^B]$ be the block of vector containing the FC between *target class* and *base class* of disease model samples. Each vector is defined element-wise $\boldsymbol{R}^b = \{r_{\tilde{g}}^b\}_{\tilde{g}\in\widetilde{\mathcal{G}}_p^b}$, where $\widetilde{\mathcal{G}}_p^b$ denotes the human gene subset $\mathcal{G}_p^b$ with its equivalent gene organism symbols for the disease model. The FC $r_{\tilde{g}}^b$ are computed through Eq (2).

$$r_{\tilde{g}}^b := \begin{cases} 0 & \mathrm{p}_{\tilde{g}}^b > 0.05 \\ sign\left(\log_2 FC_{\tilde{g}}^b\right) 2^{\log_2 FC_{\tilde{g}}^b} & \mathrm{p}_{\tilde{g}}^b \leq 0.05 \end{cases} \tag{2}$$

Where $\log_2 FC_{\tilde{g}}^b$ and $p_{\tilde{g}}^b$ are the logFC of *target class* versus *base class* disease model samples and its adjusted p-value, respectively, both obtained from FindMarkers of Seurat.

We choose adaptive sparse multi-block partial least square discriminant analysis (asmbPLS-DA) as the basic model for our method, since it discriminates between multiple disease outcome groups and selects features on high-dimensional omics data using a multi-block data structure (Zhang and Datta, 2023). The detailed introduction of asmbPLS-DA is shown in Supplementary Material S1.

asmbPLS-DA is trained on human data, with $\boldsymbol{Y}$ as the response and $\underset{\sim}{C}$ as the block-variable, always centered and scaled, with k_0 being the reference class. Therefore, for each PLS component $j = 1, \ldots, J$, we obtain estimates of response q, cell ω^{super} and gene ω^b weights indicating their relevance on discriminating between classes. Figure 1 depicts the trained model. Once the asmbPLS-DA model is trained, we can then predict the response $\boldsymbol{Y}$ for the original predictor data $\underset{\sim}{C}$ and new samples. Concretely, with the original data $\underset{\sim}{C}$ and estimated cell and gene weights, the scores t^{super} and t^b are calculated at cell and gene level, respectively, with which we predict the fitted model Eq (3).

$$\widehat{y}_{ik} = \sum_{j=1}^{J} t_{ij}^{super} q_{kj}^T \tag{3}$$

For a generic human sample and class, we name the continuous response prediction $\widehat{y}_{ik}$ as the superpathway's score.

We define in Eq (4) the *superpathway reference recapitulation* as the difference in the median superpathway's score between the *target class* and *base class* samples.

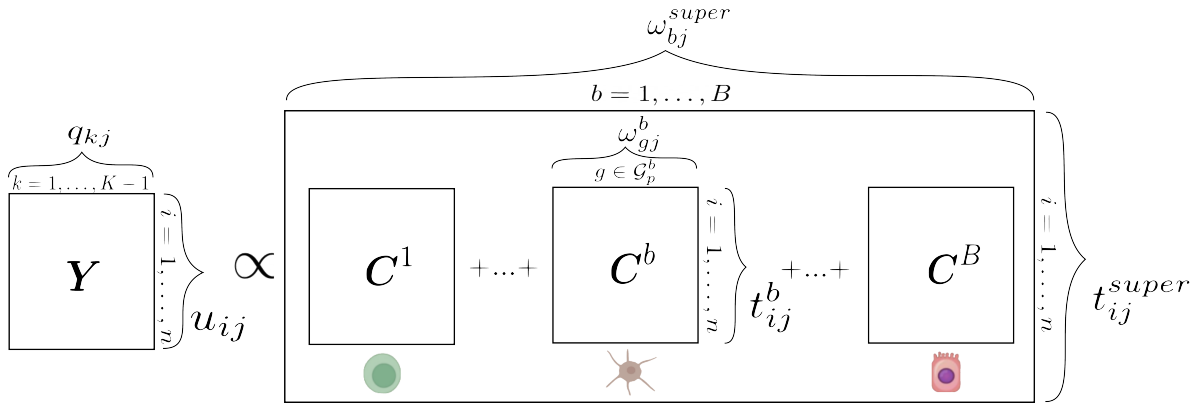

Fig. 1. Representation of asmbPLS-DA for scRNA-seq readouts. The response matrix $\boldsymbol{Y}$ contains samples as rows and classes as columns, one-hot encoded. Predictor blocks $\boldsymbol{C}^b$ are defined by cell types, with columns representing genes and rows representing samples. Each element within these predictor blocks is the pseudobulk of gene expression values. The figure displays loadings for the response matrix $\boldsymbol{q}_{kj}$, predictor blocks ω_{gj}^b and the predictor superblock ω_{bj}^{super}, as well as scores for the response matrix $\boldsymbol{u}_{ij}$, predictor blocks $\boldsymbol{t}_{ij}^b$, and the superblock $\boldsymbol{t}_{ij}^{super}$.

$$\widehat{\Omega}_{k_1} := \underset{i \in \{1 \le i \le n | y_{ik_1}=1\}}{\text{median}} (\widehat{y}_{ik}) - \underset{i \in \{1 \le i \le n | y_{ik_0}=1\}}{\text{median}} (\widehat{y}_{ik}) \quad (4)$$

In Supplementary Material S1 we show that Eq (3) can be further developed into Eq (5) and Eq (7).

$$\widehat{y}_{ik} = \sum_{b=1}^{B} \left[\sum_{j=1}^{J} t_{ij}^b \left(\omega_{bj}^{super}\right)^T q_{kj}^T \right] = \sum_{b=1}^{B} \widehat{\gamma}_{ik}^b \quad (5)$$

Where $\widehat{\gamma}_{ik}^b$ is the contribution of cell type b to the superpathway's score of class k. With the former contributions, we characterize Eq (6) as the *cell type b reference recapitulation* defined by the observed difference in median cell contribution b to superpathway's score between the *target class* and *base class*.

$$\widehat{\Gamma}_{k_1}^b := \underset{i \in \{1 \le i \le n | y_{ik_1}=1\}}{\text{median}} \left(\widehat{\gamma}_{ik}^b\right) - \underset{i \in \{1 \le i \le n | y_{ik_0}=1\}}{\text{median}} \left(\widehat{\gamma}_{ik}^b\right) \quad (6)$$

Similarly, we can extract such contributions at the gene level as shown in Eq (7).

$$\widehat{y}_{ik} = \sum_{b=1}^{B} \sum_{g \in \mathcal{G}_p^b} \left[\sum_{j=1}^{J} \frac{x_{ig}^b \omega_{gj}^b}{\sqrt{|\mathcal{G}_p^b|}} \left(\omega_{bj}^{super}\right)^T q_{kj}^T \right] = \sum_{b=1}^{B} \sum_{g \in \mathcal{G}_p^b} \widehat{\delta}_{igk}^b \quad (7)$$

Where $\widehat{\delta}_{igk}^b$ is the gene g contribution to cell type b. Note that $\widehat{\gamma}_{ik}^b = \sum_{g \in \mathcal{G}_p^b} \widehat{\delta}_{igk}^b$, hence the cell type contributions to the superpathway's score are the sum of all gene contributions within the cell type.

With the reference recapitulations, we have built a set of metrics based on the initial superpathway $\mathcal{P}^p$ that inform us on the similarity, at varying granularity levels, between *target class* and *base class* human samples.

Our aim now is to assess how similar the disease model is to the *target class* human single-cell gene expression. However, a direct comparison is not possible. For this reason, we define the *singIST treated samples* as the single-cell human gene expression samples we would have observed if human scRNA-seq behaved like the changes observed in the disease model. This is an assumption that IST and FIT follow to model human gene expression as a function of disease model changes measured through FC. With the disease model FC we derive the *singIST treated samples* by using Eq (8) in human samples belonging to the *base class*.

$$x_{ig}^{b\,\prime} := \ell\left(x_{ig}^b, r_{\tilde{g}}^b; \mu_g^b\right) = \begin{cases} x_{ig}^b & \neg A \vee \neg B \\ x_{ig}^b + \mu_g^b r_{\tilde{g}}^b & A \wedge B \wedge C \\ x_{ig}^b - x_{(1)g}^b & A \wedge B \wedge \neg C \end{cases} \quad (8)$$

Where μ_g^b is the gene centroid computed by asmbPLS-DA, an intuition of the *Biological link function* $\ell\left(x_{ig}^b, r_{\tilde{g}}^b; \mu_g^b\right)$ is shown in Supplementary Material S1. Three biologically plausible scenarios define the transformation in Eq (8); $(\neg A \vee \neg B)$ the case where either cell type b does not exist in the disease model (not A) or disease model does not have a one-to-one ortholog of human gene g (not B); $(A \wedge B \wedge C)$ the case where both cell type b and a one-to-one ortholog gene of g exist in disease model, and the translation of FC does not produce a negative expression $C = \min_{i \in \{1 \le i \le n | y_{ik_0}=1\}} \{x_{ig}^b + \mu_g^b r_{\tilde{g}}^b\} \ge 0$; $(A \wedge B \wedge \neg C)$ the case when both cell type b and one-to-one ortholog gene of g exist in the disease model, and the translation of FC does produce a negative expression any human sample (not C).

With the new predictor block $\underset{\sim}{C}' = [\boldsymbol{C}^{1\prime}, \ldots, \boldsymbol{C}^{b\prime}, \ldots, \boldsymbol{C}^{B\prime}]$ we use Eq (3) to predict the continuous response $\widehat{\boldsymbol{Y}}'$. Equally as before, we can define the *predicted recapitulations* of the disease model.

The *superpathway predicted recapitulation*.

$$\widehat{\Omega}'_{k_1} := \underset{i \in \{1 \le i \le n | y_{ik_1}=1\}}{\text{median}} (\widehat{y}'_{ik}) - \underset{i \in \{1 \le i \le n | y_{ik_0}=1\}}{\text{median}} (\widehat{y}_{ik}) \quad (9)$$

Likewise, the *cell type b predicted recapitulation*.

$$\widehat{\Gamma}_{k_1}^{b\,\prime} := \underset{i \in \{1 \le i \le n | y_{ik_1}=1\}}{\text{median}} \left(\widehat{\gamma}_{ik}^{b\,\prime}\right) - \underset{i \in \{1 \le i \le n | y_{ik_0}=1\}}{\text{median}} \left(\widehat{\gamma}_{ik}^b\right) \quad (10)$$

To ease interpretation on the similarity between human and disease model, one can express the predicted recapitulation as a fraction of the reference recapitulation, at both granularity levels: superpathway $\frac{\widehat{\Omega}'_{k_1}}{\widehat{\Omega}_{k_1}}$ and cell type b $\frac{\widehat{\Gamma}_{k_1}^{b\,\prime}}{\widehat{\Gamma}_{k_1}^b}$. In Supplementary Material S1 we prove the former fractions depend on: direction

and magnitude of FC (2), orthology $\exists\tilde{g} \equiv g$, cell type existence in disease model, and human gene ω^b and cell type weights ω^{super}.

The *superpathway predicted recapitulation* as a fraction of the *superpathway reference recapitulation.*

$$\widehat{\Omega}f_{k_1} = 100 \cdot \frac{\widehat{\Omega}'_{k_1}}{\widehat{\Omega}_{k_1}} \tag{11}$$

Cell type b predicted recapitulation as a fraction of *cell type b reference recapitulation.*

$$\widehat{\Gamma}f^b_{k_1} = 100 \cdot \frac{\widehat{\Gamma}^b_{k_1}{}'}{\widehat{\Gamma}^b_{k_1}} \tag{12}$$

To allow for downstream analysis on what drives cell type recapitulation, we define the contribution of gene $g \in \mathcal{G}^b_p$ on recapitulation (12).

$$\widehat{\Delta}f^b_{gk_1} = 100 \cdot \frac{\widehat{\Delta}^b_{gk_1}}{\widehat{\Gamma}^b_{k_1}} \tag{13}$$

Where $\widehat{\Delta}^b_{gk_1} := \widehat{\delta}^b_{igk_1}{}' - \widehat{\delta}^b_{igk_0}$, which is constant for all samples i, this fact is proven in Supplementary Material S1. Interpretation of (13) is the contribution of gene g toward *cell reference recapitulation* $\widehat{\Gamma}f^b_{k_1}$, this is due to the relationship between (12) and (13) as $\widehat{\Gamma}f^b_{k_1} = \sum_{g\in\mathcal{G}^b_p}\widehat{\Delta}f^b_{gk_1}$ (proof in Supplementary Material S1).

A graphical summary of singIST procedure is shown in Fig 2.

2.2.2 Interpretation of recapitulation measures

Recapitulation measures have a related interpretation to that of IST (Picart-Armada et al., 2024). A recapitulation of $\widehat{\Omega}f_{k_1} \approx 100$ would imply that the median of the predicted superpathway's score of singIST treated samples corresponds to that of the *target class*, hence human data show similarity to disease model. If cell recapitulation $\widehat{\Gamma}f^b_{k_1} \approx 100$ then the median of the predicted cell contribution to superpathway's score of samples singIST treated samples corresponds to that of the *target class*, and human data show similarity to disease model for cell type b. Gene contribution $\widehat{\Delta}f^b_{gk_1}$ to cell type recapitulation b provides with the magnitude of change and direction such gene has contributed to $\widehat{\Gamma}f^b_{k_1}$. This contributions will vary greatly depending on the case; positive gene contributions arise from large $r^b_{\tilde{g}}$ in disease model and agreement in direction of change in the asmbPLS-DA model in human data; negative contributions arise from opposed directions of change between $r^b_{\tilde{g}}$ and that estimated in the asmbPLS-DA with human data. On the other hand, if a cell type is not present in disease model its cell recapitulation $\widehat{\Gamma}f^b_{k_1} = 0$, and consequently all its gene contributions are null. Similarly, if gene a does not have a one-to-one disease model ortholog $\tilde{g} \equiv g$ then $\widehat{\Delta}f^b_{gk_1} = 0$. Genes having no relevance in asmbPLS-DA and/or null differences in disease model would have a recapitulation $\widehat{\Delta}f^b_{gk_1} = 0$. Details on these facts are shown in Supplementary Material S1.

2.3 Validation methodology

2.3.1 Validity test of the optimal asmbPLS-DA

Once the optimal model is selected, the validity of such for classifying between k_1 and k_0 is checked by a permutation test (Brandolini-Bunlon et al., 2019) adapted to small sample size and asmbPLS-DA. A null model distribution $H_0 : Y \perp\!\!\!\perp \underset{\sim}{C}$ is generated by permuting Y, noted as $\sigma(Y)$, and setting the number of permutations. For each $\sigma(Y)$ an asmbPLS-DA model is fitted with J^* and λ^b_j as the optimal number of PLS components and quantile combination for each block and PLS component, respectively, and randomly taking one sample out to avoid overfitting. With the permuted model, Y is predicted for all samples under analysis, such prediction is compared against the true Y by F_1 score. The rationale behind randomly permuting each Y element is that its original relationship of the model is disrupted while the dependence structure of $\underset{\sim}{C}$ is preserved (Winkler et al., 2015), thus providing a control of a false positive model. If the optimal model is actually significant it is expected that error measures increase substantially when permuting.

To this end, the F_1 LOOCV error of optimal model is compared against the $CI_{(1-\alpha)}$ of null distribution of F_1 score, where $\alpha = 0.05$ is the confidence treshold and the quantile serves as the p-value which is adjusted for multiple comparison by Benjamini-Hochberg (Benjamini and Hochberg, 1995), FDR is set to 0.1.

2.5.2 Parameters variabilities and significance of the optimal asmbPLS-DA model

Cell type and gene importance, within a cell type b, may be assessed by considering the weighted average of its estimated coefficient, by taking as weights the relative importance of each PLS component $q_{k_1 j}$ (Bougeard et al., 2011). To this end, we define the Cell Importance Projection (CIP) for cell type b:

$$CIP^b = \frac{\sum_{j=1}^{J^*} q_{k_1 j}\left(\omega^{super}_{bj}\right)^2}{\sum_{j=1}^{J^*} q_{k_1 j}} \tag{14}$$

Similarly we define the Gene Importance Projection (GIP) for gene g within cell type b:

$$GIP^b_g = \frac{\sum_{j=1}^{J^*} q_{k_1 j}\left(\omega^{b}_{gj}\right)^2}{\sum_{j=1}^{J^*} q_{k_1 j}} \tag{15}$$

Both indices verify $\sum_{b=1}^{B} CIP^b = \sum_{g\in\mathcal{G}^b_p} GIP^b_g = 1$, this is proven in Supplementary Material S1. The direction of CIP^b may be assesed by $sign(CIP^b) = sign\left(\sum_{j=1}^{J^*} q_{k_1 j}\omega^{super}_{bj}\right)$, and equivalently for GIP^b_g. Note that CIP^b distribution is nested to the already estimated $\lambda^b_j \in [0,1]$, the blocks with only small number of relevant genes will assign higher λ^b_j values, being a cell type $\lambda^b_j = 1$ a cell type that does not contain any relevant information in classifying between k_1 and k_0.

The GIP^b_g distribution of a gene that is significant is likely to substantially differ from its associated null H_0 distribution. A null distribution of GIP^b_g of the form $H_0 : x^b_{i_1 g_1} \perp\!\!\!\perp x^b_{i_2 g_2}, \forall i_1 \neq i_2, \forall g_1 \neq g_2$, and $b \in \{1,\ldots,B\}$ is generated by permuting all samples and genes within block. Note that permuting between block would not satisfy exchangeability assumption as GIP^b_g distribution is dependent on λ^b. The median $GIP^b_g \sim \mathcal{D}^b_{g,Jackknife}$ of distribution of the optimal model is compared against the null distribution $GIP^b_g \sim \mathcal{D}^b_{H_0}$ by a Mann-Whitney U test, where the alternative hypothesis being the median $\mathcal{D}^b_{g,Jackknife}$ greater than median of $\mathcal{D}^b_{H_0}$. P-value

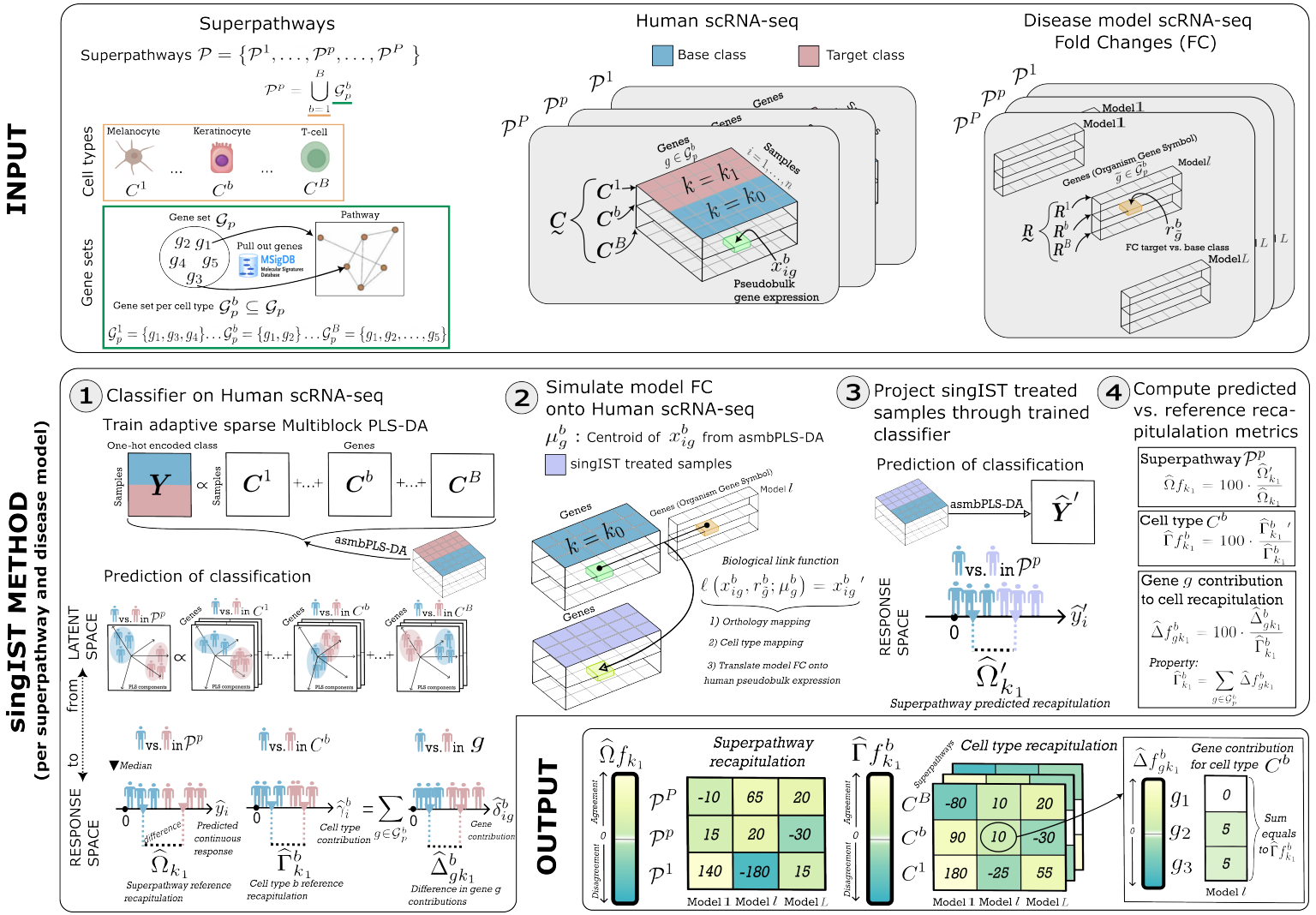

Fig. 2. INPUT) First, definition of a superpathway ($\mathcal{P}^p$) as a set containing cell types and genes. For each $\mathcal{P}^p$, there is a gene set $\mathcal{G}_p$ from which gene subsets are derived for cell types $\{\mathcal{G}_p^b\}_b \subseteq \mathcal{G}_p$. Second, for each $\mathcal{P}^p$ human scRNA-seq data is organized into matrix layers. Target class ($k = k_1$) is the human experimental group that the disease model aims to mimick (i.e disease, treated), while base class ($k = k_0$) is such that should differentiate from target class (i.e healthy control, untreated). Third, for each $\mathcal{P}^p$ disease models scRNA-seq FC are structured into vector layers. **singIST METHOD)** The method is organized into four steps, which runs independently for each $\mathcal{P}^p$ and disease model. **Step 1)** *Objective:* Quantify differences between target and base classes human samples at various levels of granularity (superpathway, cell type, and gene) using asmbPLS-DA. *Input:* A $\mathcal{P}^p$ and human scRNA-seq data. *Output:* Optimal asmbPLS-DA. From such, we derive cell type contributions ($\widehat{\gamma}_i^b$) and gene contribution ($\widehat{\delta}_{ig}^b$). With the contributions we compute similarity measures at the superpathway ($\widehat{\Omega}_{k_1}$) and the cell type levels ($\widehat{\Gamma}_{k_1}^b$). **Step 2)** *Objective:* biologically unify the human data with the disease model data for subsequent comparison. *Input:* Human scRNA-seq base class samples and disease model scRNA-seq FC data. *Output:* Human scRNA-seq gene expression observed when disease model FC are applied, we call them *singIST treated samples*. The former is achieved in the *Biological link function*, which performs steps; one-to-one orthologous mapping; cell type alignment; translate FC to x_{ig}^b, μ_g^b is a scaling factor. **Step 3)** *Objective:* Compute metrics of output from Step 1 between singIST treated samples and human scRNA-seq base class. *Input:* singIST treated samples, Human scRNA-seq base class samples and optimal asmbPLS-DA. *Output:* Pathway predicted recapitulation ($\widehat{\Omega}'_{k_1}$), Cell type b predicted recapitulation ($\widehat{\Gamma}^b{}'_{k_1}$) and predicted gene contributions ($\widehat{\delta}^b{}'_{ig}$). **Step 4)** *Objective:* Compute similarity metrics between human and disease model. *Input:* From step 1; $\widehat{\Omega}_{k_1}$ and $\widehat{\Gamma}_{k_1}^b$, and $\widehat{\delta}_{ig}^b$. From step 3; $\widehat{\Omega}'_{k_1}$ and $\widehat{\Gamma}^b{}'_{k_1}$, and $\widehat{\delta}^b{}'_{ig}$. *Output:* Predicted recapitulations as a fraction of the reference recapitulations ($\widehat{\Omega} f_{k_1}$, $\widehat{\Gamma} f_{k_1}^b$). $\widehat{\Gamma} f_{k_1}^b$ is explained by contributing genes ($\widehat{\Delta} f_{gk^*}^b$) , providing interpretation on which genes drive the cell type recapitulation. **OUTPUT)** $\widehat{\Omega} f_{k_1}$ and $\widehat{\Gamma} f_{k_1}^b$ are displayed. Positive values show agreement in gene expression change between disease model and humans, negative show opposed one. Each $\widehat{\Gamma} f_{k_1}^b$ equals to the sum of its gene contributions $\widehat{\Delta} f_{gk_1}^b$.

is adjusted by Bonferroni correction with a lower bound of expected number of true null hypothesis for each cell type $m_0^b = \left\lfloor \prod_{j=1}^{J^*} \lambda_j^b |\mathcal{G}_p^b| \right\rfloor$, the rationale is provided in Supplementary Material S1.

3. Results

All tables and figures show only superpathways under discussion, one can find tables and figures with the full 22 superpathways in Supplementary Material S3. Table 2 provides a summary of characteristics of the optimal trained models, gene set size per cell type $\mathcal{G}_p^b$ was set equal to $\mathcal{G}_p$ for all pathways. The gene set sizes of pathways are highly variable, ranging from a minimum of 15 genes for CD40/CD40L signaling [PID] to a maximum of 701 genes for Cytokine signaling to the Immune system [REACTOME]. All optimized asmbPLS-DA demonstrated statistical significance at a false discovery rate (FDR) treshold of 0.1. The optimal quantile sparsity values for Dendritic Cells in Th1/Th2 Development [BIOCARTA] show that T-cell (TC) are the mayor source of information with $\lambda^b = 0.35$ and 4 statistically significant genes, followed by Melanocytes (MC) with $\lambda^b = 0.45$ and 4 statistically significant genes, and Dendritic Cell (DC) $\lambda^b = 0.55$ and 3 statistically significant genes. On the contrary, both Keratinocyte (KC) and Langerhans Cell (LC) show the least importance with $\lambda^b = 0.95$ and 1 statistically significant gene. For JAK-STAT signaling pathway [KEGG], all cell types show the same amount of relevant information with $\lambda^b = 0.05$, suggesting an active role of all cell types in activating this pathway, with the number of statistically significant genes ranging from a minimum of 32 for TC and LC, and a maximum of 41 genes for MC. In Chemokine receptors bind chemokines [REACTOME], KC and DC were the most informative cell types with both showing a $\lambda^b = 0.05$ and 19 and 13 statistically significant genes, respectively. Lastly, Antigen Presenting Cells (APCs) were the most informative in Cytokine-Cytokine receptor interaction [KEGG] with 51 statistically significant genes for LC and 31 for DC.

To offer an overview of the most predictive genes for classifying between AD lesional and HC, Table 3 lists the five statistically significant genes with the highest GIP_g^b values, along with their corresponding directional references from the literature. In Dendritic Cells in Th1/Th2 Development, it is genes within TC that drive the prediction, concretely, IL13 shows the strongest upregulation, along with IL5 and CSF2/GM-CSF, while TLR7 shows a downregulation. ANP/CD13 shows a strong upregulation in both APCs, while for MC it is an upregulation of IL10 that drives the prediction, as well as the upregulation of antigen coding genes: ITGAX/CD11c, CD7 and CD33. All cell types in JAK-STAT present a similar importance, in line with the λ^b obtained, however top five genes within each cell type differ. In TC higher expression in AD samples compared to HC of interleukins dominate IL13, IL26, IL2RA, IL7, as well as IFNGR1. MC from AD samples show a simultaneous upregulation of CCND3 and CCND1, and so do receptors IFNGR2 and IL10RA. KC are characterized by overexpression of IL15 and its receptor IL15RA, on the contrary we observe a downregulation of SPRIY1. Further, IL23A and IFNL1/IL-29 are downregulated in DC, while SPRED1, SOCS1 and OSM are upregulated. For LC a downregulation of genes IL22RA2, JAK1 and CCND1 is estimated, on the contrary there is a downregulation of CCND2 and STAT6. APCs are the mayor drivers of Cytokine-Cytokine receptor interaction pathway predictivity. Precisely, for DC chemokine receptor and ligands CCR6, CCL5/RANTES and CCL3L1 were all supressed in AD, and Herpesvirus entry mediator gene TNFRSF14 was upregulated. IL23A is inhibited in both LC and DC. Chemokine receptors bind chemokines pathway are fueled by KC, DC and LC. Common chemokine AD markers are observed; CCL20/LARC, CCL7/MCP1 and CCR2 upregulatione in KC; upregulated CCL17/TARC, CRR1 and CXCR4 in LC; CCL13 in DC.

One-to-one orthology of each disease model against humans was assessed, likewise their superpathway recapitulations from singIST procedure are shown in Figure 3. All disease models exhibit a similar level of observed one-to-one orthology, with OVA typically displaying fewer sequenced genes per pathway. However, orthology coverage varies significantly by pathway. Notably, the Asthma and Chemokine receptors bind chemokines pathways show 60% orthology coverage, while the IL12 signaling mediated by STAT4 pathway exhibits 100% coverage. Superpathway recapitulations for Dendritic Cells in Th1/Th2 Development show overall agreement in direction with human AD for OXA (96.6%), as the highest recapitulation, followed by IMQ (41.4%), and OVA (0%) showing no recapitulation at all. Cell type recapitulations are displayed in Figure 4, which show high TC recapitulation for IMQ (138.1%) and no recapitulation for OVA (0%) and OXA (1.3%). Concretely, OXA is driven by LC (643.7%) and DC (253.1%) recapitulations. OVA has 0% recapitulation for LC across all pathways, since the LC cell type had fewer than 100 cells in OVA, they were removed. MC have 0% recapitulation across all disease models and pathways. Figure 5 illustrates the gene contribution to cell type recapitulation. TC recapitulation of IMQ come from CSF2/GM-CSF (129.8%) and IL5 (46.4%), while TLR7 (-39.7%) shows disagreement with human condition, since its FC is 6.1 while TLR7 is suppressed in human AD. OXA recapitulation is solely driven by ANPEP with strong FC in LC (13.3) and DC (2.8), which aligns with the observed upregulation of ANPEP in human APCs. OVA stands out for having no DEGs in this pathway, resulting in 0% recapitulation. Interestingly, strongest TC marker IL13 in AD is not DEG by any disease model. In the JAK-STAT signaling superpathway recapitulations, OXA (84.3%) performs the highest while IMQ (13.9%) and OVA (1.5%) show low recapitulations. Precisely, OXA agrees in direction with moderate recapitulations for all cell types, strongest contributing genes are IL21R (-31.5%) in DC, OSM (58.5%; 118.8%; 66.1%) across APCs and KC, and all STAT1/2/3/5 in TC. Gene contributions in TC for IMQ generally agree with OXA with disagreement in direction and/or magnitude on: CSF2 (16.6%), CSF3R (-1.8%), IL23R (-1%), IL2RG (-2.2%) and JAK1 (-5.6%). Cytokine-Cytokine receptor interaction pathway show no superpathway recapitulation at all for IMQ (0.6%) and OVA (0.6%), while OXA (380%) a very extreme recapitulation. For IMQ and OVA, cell type recapitulation is very heterogeneous with some agreeing on direction (TC, LC) and others disagreeing (KC, DC), which overall drives the superpathway recapitulation to almost zero. The same holds for OVA. OXA extreme recapitulation in LC (860.2%) comes from INHBA (745.5%) with an extreme FC of 627.2, while disagreement in KC (-35.5%) are mainly from OSM (-21.6%), CRR5 (-8.8%), IL23R (-5.2%). Superpathway recapitulations of Chemokine receptors bind chemokines depict high differences between the murine models, high recapitulation for OVA (105.2%), disagreement in direction for IMQ (-25.5%) and low recapitulation for OXA (7.5%). OVA recapitulation comes purely from KC (317.7%) with CXCL6 (312.6%) being the

Table 2. Characteristics of fitted optimal asmbPLS-DA for all superpathways.

Pathway	J^{*a}	Gene set size^b	adj pvalue^c	Dendritic Cell^d		Keratinocyte		Langerhans Cell		Melanocyte		T-cell	
				λ^1	GIP_g^1*	λ^2	GIP_g^2*	λ^3	GIP_g^3*	λ^4	GIP_g^4*	λ^5	GIP_g^5*
Cytokine-Cytokine receptor interaction [KEGG]	2	263	$p \leq 0.001$	0.75 0.05	31	0.95 0.05	17	0.55 0.05	51	0.95 0.65	29	0.95 0.45	15
Chemokine receptors bind chemokines [REACTOME]	1	57	$p \leq 0.001$	0.05	13	0.05	19	0.45	9	0.95	3	0.85	6
Dendritic Cells in Th1/Th2 Development [BIOCARTA]	1	17	$p \leq 0.05$	0.55	3	0.95	1	0.95	1	0.45	4	0.35	4
JAK-STAT signaling pathway [KEGG]	1	155	$p \leq 0.1$	0.05	35	0.05	36	0.05	32	0.05	41	0.05	32

a. Optimal number of PLS components, for LOOCV was set to $J \leq 3$, *b.* Gene set size *c.* Adj. p-val of global significance, *d.* λ^b optimal quantiles of cell type b. The quantile space were set to 100000 combinations of λ^b values ranging along $\{0.05, \ldots, 0.55, \ldots, 0.95\}$. GIP_g^b* number of statistically significant GIP_g^b whose $adj.p - value \leq 0.05$, permutation tests were run on 10000 permutations.

Table 3. Top five statistically significant genes ordered by GIP_g^b magnitude.

Pathway	Cell type	CIP (direction)	Top 5 genes^a	GIP (direction)	Reference direction^{b,c,d}
JAK-STAT signaling pathway [KEGG]	T-cell	0.17 (↑)	IL13, IL26, IL2RA, IL7, IFNGR1	0.05 (↑), 0.03 (↑), 0.03 (↑), 0.03 (↑), 0.03 (↑)	↑, ↑, ↑, −, −
	Dendritic Cell	0.19 (↑)	IL23A, SPRED1,SOCS1, IFNL1, OSM	0.05 (↓), 0.04 (↑), 0.04 (↑), 0.03(↓), 0.03 (↑),	↑†,↑*,†,↑†,↓†,↑
	Langerhans Cell	0.17 (↑)	IL22RA2, CCND2, CCND1, JAK1, STAT6	0.04 (↑), 0.04 (↓), 0.04 (↑), 0.04 (↑), 0.03 (↓)	↑,−,−, ↑†,−
	Keratinocyte	0.28 (↑)	CCND3, SPRY1, IL15RA, IFNAR2, IL15	0.03 (↑), 0.03 (↓), 0.02 (↑), 0.02 (↑), 0.02 (↑)	−,↓*, ↑*,↑*,↑
	Melanocyte	0.18 (↑)	CCND3, IFNGR2, CCND2, IL10RA, CCND1	0.03 (↑), 0.03 (↑), 0.03 (↑), 0.03 (↑), 0.03(↑)	↑*,↑*,−, ↑*,↑*
Dendritic Cells in Th1/Th2 development [BIOCARTA]	T-cell	0.38 (↑)	IL13, IL5, CSF2, TLR7	0.42 (↑), 0.13 (↑), 0.11 (↑), 0.11 (↓)	↑, ↑, ↑, ↓*
	Dendritic Cell	0.16 (↑)	ANPEP, CSF2, IL13	0.85 (↑), 0.06 (↑), 0.04 (↑)	↑†, ↑†, ↑
	Langerhans Cell	0.11 (↑)	ANPEP	1 (↑)	↑†
	Keratinocyte	0.09 (↑)	ITGAX	1 (↑)	↑†
	Melanocyte	0.26 (↑)	IL10, ITGAX, CD7, CD33	0.39 (↑), 0.23 (↑), 0.14 (↑), 0.10 (↑)	↑†, ↑†, −, −
Cytokine-Cytokine receptor interaction [KEGG]	T-cell	0.11 (↑)	IL13, CCR2, TNFSF10, CXCL13, IL1R2	0.28 (↑), 0.22 (↓), 0.12 (↑), 0.09 (↑), 0.01 (↑)	↑, ↓, ↑, ↑†, ↑
	Dendritic Cell	0.32 (↑)	IL23A, CCR6, CCL5, CCL3L1, TNFRSF14	0.11 (↓), 0.06 (↓), 0.06 (↓), 0.05 (↓), 0.05 (↑)	↑†,↓,↓e,†, ↓†,↑
	Langerhans Cell	0.36 (↑)	IL22RA2, PLEKHO2, IL23A, IL7R, CCR1	0.06 (↑), 0.04 (↑), 0.04 (↓), 0.03 (↑), 0.03 (↑)	↑, ↑†, ↓†, ↑, ↑
	Keratinocyte	0.11 (↑)	IL15RA, TNFRSF12A, CCR2, TNFRSF11A, IFNAR2	0.41 (↑), 0.17 (↑), 0.12 (↑), 0.04 (↓), 0.05 (↑)	↑*, ↑, ↑, −, ↑*
	Melanocyte	0.10 (↑)	CX3CL1, TGFB2, IFNGR2, TNFSF13B, IL10RA	0.32 (↓), 0.28 (↓), 0.08 (↑), 0.08 (↑), 0.05 (↑)	↓*,†,↓†,↑*,−,↑*
Chemokine receptors bind chemokines [REACTOME]	T-cell	0.11 (↑)	CCR2, CXCL13, CCR1, CXCL8, CCR7	0.42 (↓), 0.31 (↑), 0.12 (↑), 0.07 (↑), 0.03 (↓)	↓, ↑†, ↑, ↑ ↓
	Dendritic Cell	0.32 (↑)	CCL5, CCR6, CCL3L1, CCL13, CXCL2	0.07 (↓), 0.07 (↓), 0.07 (↓), 0.06 (↑), 0.06 (↓)	↓e,†,↓, ↓†,↑†,−
	Langerhans Cell	0.17 (↑)	CCR1, CCL17, CXCR4, CCR10, CCRL2	0.17 (↑), 0.13 (↑), 0.12 (↑), 0.10 (↓), 0.08 (↑)	↑, ↑, ↑, −, ↑†
	Keratinocyte	0.36 (↑)	CCR2, CXCL2, CCL7, XCR1, CCL20	0.07 (↑), 0.06 (↑), 0.05 (↑), 0.04 (↑), 0.04 (↑)	↑, −, ↑†, − , ↑
	Melanocyte	0.04 (↑)	CX3CL1, CXCL8, CXCR6	0.95 (↓), 0.03 (↑), 0.02 (↑)	↓*,†, ↑†, ↑†

a. Only genes with $FDR \leq 0.05$ and maximum GIP_g^b magnitude. *b.* ↑ for upregulated genes in literature, ↓ for supressed and − for unknown/inconsistent direction. *c.* References on direction are in Supplementary Material S3. *d.* ∗: not reported in human skin AD but reported gene's deregulation effect in specific cell type, disease models or human AD PBMC. †: reported in bulk but not cell type. *e.* in chronic AD skin lesions.

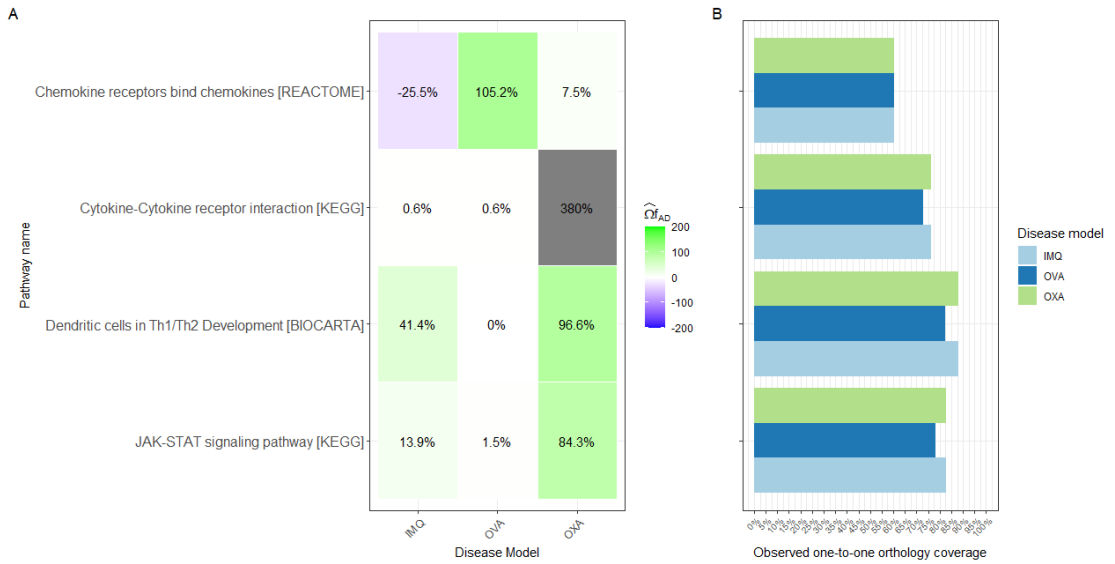

Fig. 3. Superpathway recapitulation and observed one-to-one orthology of AD disease models. A) Superpathway predicted recapitulation as a fraction of the superpathway reference recapitulation for IMQ, OXA and OVA across all pathways under study. Negative recapitulations refer to opposed directions with human observed condition, while positive recapitulations define agreement in direction. **B)** Observed one-to-one orthology coverage refers to number of observed and one-to-one ortholog genes in disease model as a fraction of pathway gene set size. Despite all disease models belong to the same organism *mus musculus* their differences in observed orthology one-to-one coverage come from sequenced reads.

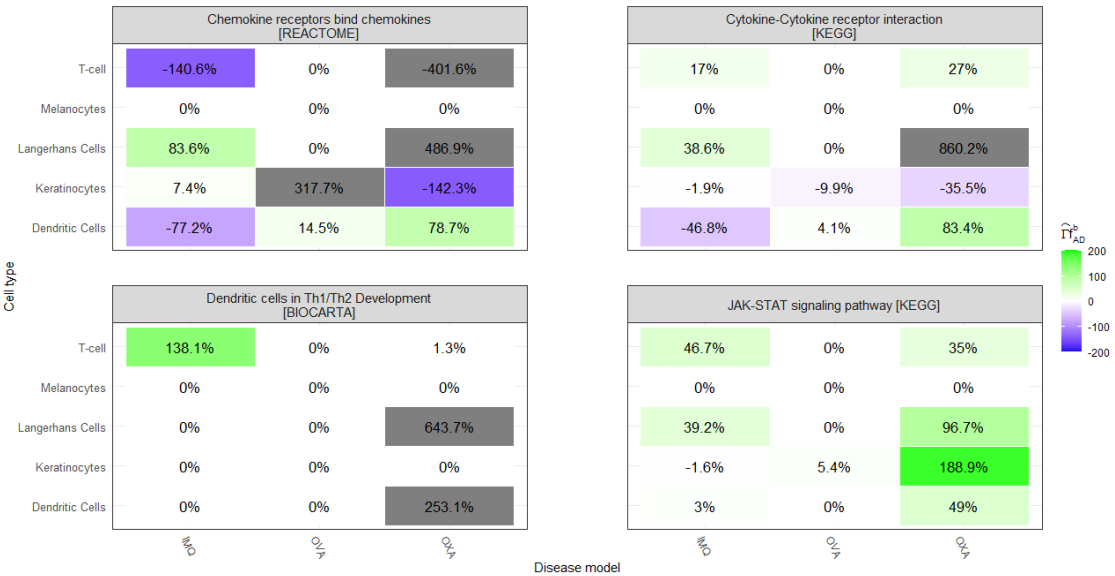

Fig. 4. Cell type recapitulation for all AD disease models and superpathways under analysis.

mayor source of contribution due to its extreme FC (499.6) and agreement with human direction. On the contrary, OXA and IMQ have higher variability between cell type recapitulations, while they agree on LC recapitulation direction, they disagree on DC and KC with IMQ (-77.2%; 7.4%) and OXA (79.7%;-142.3%). The source of differences for DC are CCL5 (IMQ -92.5%, OXA 0%), CCR5 (IMQ 8.5%; OXA 45.5%). While both IMQ and OXA show opposite directions to that of humans in TC, both mainly due to disagreement in CCL5 and CCR2.

4. Discussion

Atopic Dermatitis (AD) represents a chronic skin-immune-mediated inflammatory disease (IMID), characterized by dysregulated T-cell mediated inflammation and keratinocyte differentiation (Tsoi et al., 2019). We put a special focus in the discussion on pathways proven to be causal drivers of AD pathogenesis or related to its clinical severity; JAK-STAT signaling pathway, Dendritic Cells in regulating Th1/Th2 development, Cytokine-Cytokine receptor interaction, and Chemokine Signaling pathway. As anticipated, distinct cell types predominantly activate each pathway: TC and DC in

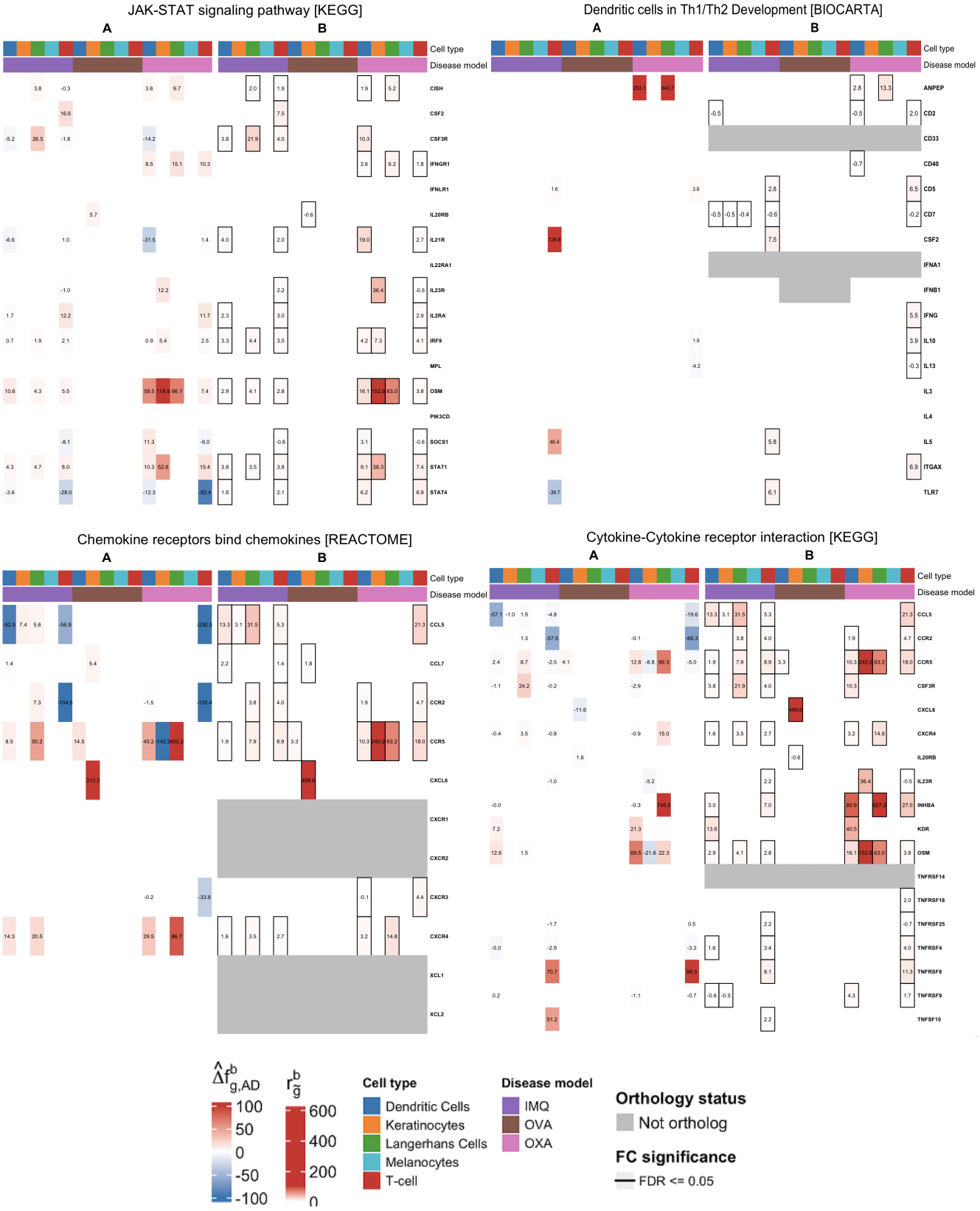

Fig. 5. Gene contribution and disease model estimated $r^b_{\tilde{g}}$. A) Gene contribution to cell type recapitulation by disease model. If gene set size of pathway is greater than 50, only the top 5 contributing genes, for each cell type, were displayed. Blank gene contributions correspond to 0 values. **B)** Computed $r^b_{\tilde{g}}$ by disease model. Grey FC refer to genes without one-to-one ortholog and/or not sequenced in disease model. Framed FC refer to statistically significant $FDR \leq 0.05$ genes, as per FindMarkers. Blank FC correspond to 0 values.

Dendritic Cells in Th1/Th2 Development, KC and DC in Chemokine receptors bind chemokines, LC and DC in Cytokine-Cytokine receptor interaction. Conversely, in the JAK-STAT signaling pathway, all cell types play a significant role in its activation. The JAK-STAT pathway is involved in a variety of biological functions beyond immunity, including but not limited to: cell division, cell death, and tumor formation. AD is characterized by intense itching and scratching, which leads to eczematous lesions. These lesions damage MC and KC which in turn activate the JAK-STAT signaling pathway, promoting cell-cycle and inflammation. It is reported that simultaneous activation of CCND3/Cyclin-D3 and CCND1/Cyclin-D1 in melanocytic skin lesions promotes cell-cycle regulation (Alekseenko et al., 2010). Additionally, MYC is upregulated and involved in this function. Furthermore, unique genes in the PIK3T-AKT subpathway, such as PIK3CD and PIK3CB, which are involved in cell-cycle and cell survival functions, are also upregulated in the JAK-STAT pathway. Similarly, upregulation in KC of IL15RA/IL15Rα and IL15 in skin lesions has been associated to inflammatory processes. Additionally, the activation of the JAK-STAT pathway by TC, LC and DC is primarily related to immune functions.

It was interesting, though not surprising, to observe the low number of DEGs in OVA for the pathways under study. This finding aligns with the low DEGs observed in TC and DC clusters reported in Leyva-Castillo et al. (2022), as well as bulk-RNA studies showing a low number of DEG in characteristic pathways of AD (Ewald et al., 2017). The absence of MC across all disease models is expected, as the interfollicular epidermis of mouse pelage skin, particularly in the ear where biopsies are collected, entirely lacks functional, pigment-producing MC (Michalak-Mićka et al., 2022).

Dendritic Cells in Th1/Th2 Development pathway varies significantly across different disease models. In the IMQ model, there is a notable upregulation of TLR7 (FC 6.1), whereas this gene is suppressed in human AD, leading to a negative gene contribution of -39.7%. Although TLR7 has not been extensively studied in AD, its role in regulating TC differentiation is well-documented. Suppression of TLR7 in TC has been shown to produce a skewed Th2 immune response (Jeisy-Scott et al., 2011), consistent with the immune response observed in AD. Conversely, upregulation of TLR7 in murine models leads to the differentiation of TC towards Th17 and/or Th1 (Ye et al., 2017). In the IMQ model, upregulation of TLR7 is expected for two reasons; first, the mechanism of action of Imiquimod cream is as a TLR7 agonist; second, IMQ is characterized by a Th17-driven immune response. However, in human AD, suppression of TLR7 would be anticipated due to its Th2-skewed response. Furthermore, it is known that none of these disease models accurately mimic the mechanisms that mediate expression of IL13/IL4 genes. While the IMQ model is characterized by an increase in IL17/IL22 and not IL13/IL4, typical of psoriatic lesions, OVA and OXA models produce IL13/IL4 through infiltrating basophils (myeloid cells) rather than Th2 cells (Leyva-Castillo et al., 2022; Liu et al., 2020), contrary to previous thought, leading to contribution of such genes of 0%, or even negative for IL13 in OXA (-4.2%).

In the JAK-STAT pathway, OXA demonstrated the best superpathway recapitulation at 84.3%, while OVA performed the worst at 1.5%. OXA is preferentially chosen as a preclinical in vivo model for testing JAK inhibitors in AD (Zhang et al., 2024). One reason for OVA's poor performance is its low number of DEGs in the JAK-STAT pathway. Previous bulk-RNA comparisons of OVA and OXA in the JAK-STAT pathway highlight the superiority of OXA (Ewald et al., 2017), underscoring the deficiency of low DEGs in OVA. In TC of OXA, large negative contributions were observed for STAT4, and moderate negative ones for DC and TC in both IMQ and OXA. STAT4 levels are greatly suppressed during physiological Th2 development (Usui et al., 2003), as observed in human AD. However, both IMQ and OXA immune responses are not characterized by Th2-skewed response, hence expression of STAT4 in DC and TC is upregulated, with strong FC in opposite direction to that estimated in human AD. A similar pattern is observed in DC for IL21R both in OXA and IMQ, while IL21R is stablished as an upregulated cytokine in skin human and murine AD lesions for TC (Jin et al., 2009), less its known about its role in DC. Both OXA and IMQ show positive contributions aligned with human AD in TC for IL21R (IMQ: 1%, OXA: 1.4%) and negative ones for DC (IMQ: -6.6%, OXA: -31.5%).

The pathways Chemokine receptors bind chemokines and Cytokine-Cytokine receptor interaction show overlapping genes with large contributions. The expression of many chemokines and cytokines are not stable throughout the lifespan of the lesion; instead, their levels fluctuate depending on the stage of it. For instance, CCL5 shows large negative contributions in both pathways across TC for both IMQ and OXA, and additionally DC for IMQ. Particularly, CCL5 expression in human skin AD lesions is upregulated in the acute phase but suppressed in chronic AD wounds (Tsoi et al., 2020), here we observe a downregulation of CCL5 in human AD. Despite the lack of clinical information on lesion stages in Bangert et al. (2021), since all patients were diagnosed with chronic AD during early childhood, it is likely that their collected lesions are chronic. However, collected skin lesions from murine models are acute, which explains the difference in direction to human AD across all disease models and cell types. In addition, it is known that IMQ broadly upregulates CCL5/CCL4, while this is clear in CCL5 since TC, DC, LC and KC, show contributions, CCL4 is not a one-to-one ortholog with humans which leads to null contributions across all cell types. On the contrary, we observe CCL5 expression in OXA is more localized to TC as shown in Liu et al. (2020). CCL5 receptor CCR5 is also highly upregulated across all disease models, while OXA showing stronger contributions across all cell types compared to IMQ in line with Liu et al. (2020).

While demonstrating significant capabilities, singIST presents several limitations, including dependence on pre-annotated cell types, the assumption of homogeneous effects when translating fold changes to human gene expression, and the requirement for well-defined human disease states (e.g., endotypes) prior to analysis. Additionally, extensions could be explored on differentiating changes due to cell type-specific gene expression and cell type proportions, between human classes. Further validation in additional disease contexts will solidify its utility in drug development and preclinical research.

6. Conclusions

Here we have developed singIST an extension of IST method for comparative single-cell transcriptomics, offering a novel, integrative framework to evaluate disease model similarity to human conditions at various biological levels. singIST provides explainable and quantitative insights into transcriptomic alignment. Its application to murine models of atopic dermatitis

demonstrated the method's ability to recover known findings and generate novel hypotheses.

Supplementary data

Supplementary Data are available online.

Conflict of interest

A.M., S.P. and F.F., were all paid employees by Almirall S.A and may hold shares in the company.

Author contribution statement

A.M conceptualized and formalized the method, implemented it and performed the data analysis, interpreted data results and drafted initial article for review. S.P conceptualized the method. S.P, A.P and F.F revised critically the work for important intellectual content and approved the final version to be published.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (www.mineco.gob.es) PID2021-122952OB-I00, DPI2017-89827-R, Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), initiatives of Instituto de Investigación Carlos III (ISCIII), and with the support of the Pla de Doctorats Industrials de la Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya. B2SLab is certified as 2017 SGR 952.

A.M would like to acknowledge; Dr. Sergio Oller-Moreno and Tomas Romero-Rodriguez for software support; Dr. Bruna Oriol-Tordera and Mercè Pont-Giralt for providing useful references in AD and murine models; Dr. Juan Luis-Trincado for scRNA-seq bioinformatics support; Dr. Estrella Lozoya-Toribio for her summary skills.

References

Alekseenko, A et al. (2010) "Cyclin D1 and D3 expression in melanocytic skin lesions". In: *Archives of Dermatological Research* 302.7, pp. 545–550.

Ali, Muhammad et al. (2024) "Single cell transcriptome analysis of the THY-Tau22 mouse model of Alzheimer's disease reveals sex-dependent dysregulations". In: *Cell Death Discovery* 10.1, p. 119. ISSN: 2058-7716.

Bangert, C. et al. (2021) "Persistence of mature dendritic cells, TH2A, and Tc2 cells characterize clinically resolved atopic dermatitis under IL-4R blockade". In: *Science Immunology* 6.55, eabe2749.

Benjamini, Yoav and Yosef Hochberg (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246.

Bougeard, Stéphanie et al. (2011) "Multiblock redundancy analysis: interpretation tools and application in epidemiology". In: *Journal of Chemometrics* 25.9, pp. 467–475.

Brandolini-Bunlon, M. et al. (2019) "Multi-block PLS discriminant analysis for the joint analysis of metabolomic and epidemiological data". In: *Metabolomics* 15.134.

Brunner, Peter M et al. (2017) "The atopic dermatitis blood signature is characterized by increases in inflammatory and cardiovascular risk proteins". In: *Scientific reports* 7.1, p. 8707.

Elitt, Matthew S et al. (May 2018) "Drug screening for human genetic diseases using iPSC models". In: *Human Molecular Genetics* 27.R2, R89–R98. ISSN: 0964-6906.

Emmerich, Christoph H. et al. (2021) "Improving target assessment in biomedical research: the GOT-IT recommendations". In: *Nature Reviews Drug Discovery* 20.1, pp. 64–81.

Ewald, David A. et al. (2017) "Major differences between human atopic dermatitis and murine models, as determined by using global transcriptomic profiling". In: *The Journal of Allergy and Clinical Immunology* 139.2, pp. 562–571.

Franzén, Lovisa et al. (2024) "Mapping spatially resolved transcriptomes in human and mouse pulmonary fibrosis". In: *Nature Genetics*. ISSN: 1546-1718.

Gao, Shouguo et al. (2021) "Comparative Transcriptomic Analysis of the Hematopoietic System between Human and Mouse by Single Cell RNA Sequencing". In: *Cells* 10.5. ISSN: 2073-4409.

Gillespie, M. et al. (2022) "The Reactome Pathway Knowledgebase 2022". In: *Nucleic Acids Research* 50.D1, pp. D687–D692.

Jeisy-Scott, V. et al. (2011) "Increased MDSC accumulation and Th2 biased response to influenza A virus infection in the absence of TLR7 in mice". In: *PloS One* 6.9, e25242.

Jin, H. et al. (2009) "IL-21R is essential for epicutaneous sensitization and allergic skin inflammation in humans and mice". In: *The Journal of clinical investigation* 119.1, pp. 47–60.

Kanehisa, Minoru et al. (2023) "KEGG for taxonomy-based analysis of pathways and genomes". In: *Nucleic acids research* 51.D1, pp. D587–D592.

Karmele, EP et al. (2023) "Single cell RNA-sequencing profiling to improve the translation between human IBD and in vivo models". In: *Front Immunol* 14. PMID: 38179052; PMCID: PMC10766350, p. 1291990.

Kim, Doyoung et al. (2019) "Research Techniques Made Simple: Mouse Models of Atopic Dermatitis". In: *Journal of Investigative Dermatology* 139.5, 984–990.e1. ISSN: 0022-202X

Lawhorn, CM et al. (2018) "Simple Comparative Analyses of Differentially Expressed Gene Lists May Overestimate Gene Overlap". In: *J Comput Biol* 25.6. PMID: 29658777; PMCID: PMC5998827, pp. 606–612.

Leyva-Castillo, J. M. et al. (2022) "Single-cell transcriptome profile of mouse skin undergoing antigen-driven allergic inflammation recapitulates findings in atopic dermatitis skin lesions". In: *The Journal of Allergy and Clinical Immunology* 150.2, pp. 373–384.

Li, J et al. (2023) "Single-cell transcriptome dataset of human and mouse in vitro adipogenesis models". In: *Sci Data* 10.1. PMID: 37328521; PMCID: PMC10275883, p. 387.

Liberzon, Arthur et al. (2015) "The Molecular Signatures Database (MSigDB) hallmark gene set collection". In: *Cell systems* 1.6, pp. 417–425.

Liu, Y. et al. (2020) "Single-Cell Profiling Reveals Divergent, Globally Patterned Immune Responses in Murine Skin Inflammation". In: *iScience* 23.10, p. 101582.

Loewa, Anna et al. (2023) "Human disease models in drug development". In: *Nature Reviews Bioengineering* 1.8, pp. 545–559.

McInnes, Iain B. and Ellen M. Gravallese (Oct. 2021) "Immune-mediated inflammatory disease therapeutics: past, present and future". In: *Nature Reviews Immunology* 21.10, pp. 680–686. ISSN: 1474-1741.

Michalak-Mićka, K. et al. (2022) "Characterization of a melanocyte progenitor population in human interfollicular epidermis". In: *Cell Reports* 38.9, p. 110419.

Nishimura, Darryl (2001) "BioCarta". In: *Biotech Software & Internet Report* 2.3, pp. 117–120.

Normand, Rachelly et al. (2018) "Found In Translation: a machine learning model for mouse-to-human inference". In: *Nature Methods* 15.12, pp. 1067–1073.

Picart-Armada, Sergio et al. (2024) "In Silico Treatment: a computational framework for animal model selection and drug assessment". In: *bioRxiv*.

Pisetsky, David S. (2023) "Pathogenesis of autoimmune disease". In: *Nature Reviews Nephrology* 19.8, pp. 509–524.

Schaefer, Carl F. et al. (Oct. 2008) "PID: the Pathway Interaction Database". In: *Nucleic Acids Research* 37.suppl$_1$, pp. D674–D679. ISSN: 0305-1048.

Shegokar, Ranjita (2020) "Chapter 2 Preclinical testing Understanding the basics first". In: *Drug Delivery Aspects*. Ed. by Ranjita Shegokar. Elsevier, pp. 19–32. ISBN: 978-0-12-821222-6.

Steinmetz, Karen L. and Edward G. Spack (2009) "The basics of preclinical drug development for neurodegenerative disease indications". In: *BMC Neurology* 9.1, S2.

Storey, Joanne et al. (2022) "A Structured Approach to Optimizing Animal Model Selection for Human Translation: The Animal Model Quality Assessment". In: *ILAR Journal* 62.1-2, pp. 66–76.

Tsoi, L. C. et al. (2020) "Progression of acute-to-chronic atopic dermatitis is associated with quantitative rather than qualitative changes in cytokine responses". In: *The Journal of Allergy and Clinical Immunology* 145.5, pp. 1406–1415.

Tsoi, Lam C. et al. (2019) "Atopic Dermatitis Is an IL-13–Dominant Disease with Greater Molecular Heterogeneity Compared to Psoriasis". In: *Journal of Investigative Dermatology* 139.7, pp. 1480–1489. ISSN: 0022-202X

Usui, T. et al. (2003) "GATA-3 suppresses Th1 development by downregulation of Stat4 and not through effects on IL-12Rbeta2 chain or T-bet". In: *Immunity* 18.3, pp. 415–428.

Wei, F. et al. (2021) "A review for cell-based screening methods in drug discovery". In: *Biophysics reports* 7.6. Health Science Center, School of Pharmacy, Xi'an Jiaotong University, Xi'an 710061, China, pp. 504–516.

Winkler, Anderson M. et al. (2015) "Multi-level block permutation". In: *NeuroImage* 123, pp. 253–268. ISSN: 1053-8119.

Ye, J. et al. (2017) "TLR7 Signaling Regulates Th17 Cells and Autoimmunity: Novel Potential for Autoimmune Therapy". In: *Journal of Immunology* 199.3, pp. 941–954.

Zhang, R and S Datta (2023) "Adaptive Sparse Multi-Block PLS Discriminant Analysis: An Integrative Method for Identifying Key Biomarkers from Multi-Omics Data". In: *Genes (Basel)* 14.5, p. 961.

Zhang, Xiaotuan et al. (May 2024) "Preclinical evaluation of Janus Kinase inhibitors in atopic dermatitis: Insights from an oxazolone-induced mouse model". In: *The Journal of Immunology* 212.1$_S$*upplement*, 1415$_4$840–1415$_4$840. ISSN: 0022-1767.