Identifying maximally informative signal-aware representations of single-cell data using the Information Bottleneck

Serafima Dubnov[1,2], Zoe Piran[3], Hermona Soreq[1,2], Mor Nitzan[3,4,5,*]

1. The Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel;
2. The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel;
3. School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel;
4. Racah Institute of Physics, The Hebrew University, Jerusalem, Israel;
5. Faculty of Medicine, The Hebrew University, Jerusalem, Israel;

* Correspondence: mor.nitzan@mail.huji.ac.il; +972-2-5494686.

Classification

Biological Sciences / Biophysics and Computational Biology

Keywords

Single-cell RNA-seq, Information Bottleneck, gene clustering

Abstract

Rapid advancements in single-cell RNA-sequencing (scRNA-seq) technologies revealed the richness of myriad attributes encompassing cell identity, such as diversity of cell types, organ-of-origin, or developmental stage. However, due to the large scale of the data, obtaining an interpretable compressed representation of cellular states remains a computational challenge. For this task we introduce bioIB, a method based on the Information Bottleneck algorithm, designed to extract an optimal compressed representation of scRNA-seq data with respect to a desired biological signal, such as cell type or disease state. BioIB generates a hierarchy of weighted gene clusters, termed metagenes, that maximize the information regarding the signal of interest. Applying bioIB to a scRNA-seq atlas of differentiating macrophages and setting either the organ-of-origin or the developmental stage as the signal of interest provided two distinct signal-specific sets of metagenes that captured the attributes of the respective signal. BioIB's representation can also be used to expose specific cellular subpopulations, for example, when applied to a single-nucleus RNA-sequencing dataset of an Alzheimer's Disease mouse model, it identified a subpopulation of disease-associated astrocytes. Lastly, the hierarchical structure of metagenes revealed interconnections between the corresponding biological processes and cellular populations. We demonstrate this over hematopoiesis scRNA-seq data, where the metagene hierarchy reflects the developmental hierarchy of hematopoietic cell types.

Significance

Single-cell gene expression represents an invaluable resource, encoding multiple aspects of cellular identity. However, its high complexity poses a challenge for downstream analyses. We introduce bioIB, a methodology based on the Information Bottleneck, that compresses data while maximizing the information about a biological signal-of-interest, such as disease state. bioIB generates a hierarchy of metagenes, probabilistic gene clusters, which compress the data at gradually changing resolutions, exposing signal-related processes and informative connections between gene programs and their corresponding cellular populations. Across diverse single-cell datasets, bioIB generates distinct metagene representations of the same dataset, each maximally informative relative to a different signal; uncovers signal-associated cellular populations; and produces a metagene hierarchy that reflects the developmental hierarchy of the underlying cell types.

Main text

Introduction

Cellular gene expression profiles encapsulate a wealth of information regarding a cell's identity, defined by a variety of biological factors, such as cell type, disease state, and developmental stage. Single-cell RNA-sequencing (scRNA-seq) technologies, quantifying gene expression levels at single-cell resolution, are invaluable for revealing these facets, allowing to study the different factors encompassing a cell's identity[1]. However, exposing such factors poses a computational challenge due to the complexity and high dimensionality of scRNA-seq. While datasets typically comprise thousands of gene profiles across thousands to hundreds of thousands of cells, any reduction in dimensionality will naturally result in loss of information[2]. Specifically, when aiming to uncover factors associated with a specific biological signal (e.g. gene programs associated with disease progression), the challenge can be framed as a trade-off between reducing the complexity of the data while retaining as much information as possible regarding the signal of interest.

The Information bottleneck (IB) theory[3] allows us to reason mathematically about this trade-off. Given a dataset (e.g. scRNA-seq measurements) and a variable of interest encoded in the data (e.g. healthy vs. disease samples), IB provides a reduced data representation which is maximally informative about the selected variable[3,4]. Since it was first introduced, IB has been successfully applied in diverse fields, such as text clustering[5], image analysis[6,7], language processing[8], neuroscience[9] and computational biology[10–12].

Here, we present bioIB, a single-cell tailored method based on the IB algorithm, providing a compressed, signal-informative representation of single-cell data. The compressed representation is given by metagenes, which are clustered probabilistic mapping of genes. The probabilistic construction preserves gene-level biological interpretability, allowing characterization of each metagene.

By considering the trade-off, and focusing on the signal of interest, bioIB differs from previous approaches suggested for data compression, namely dimensionality reduction of single-cell data (e.g. PCA[13,14], ICA[15], NMF[16]), typically applied to the data without any constraints and producing factors with limited biological interpretability. Furthermore, the interpretability of bioIB's output poses an advantage compared to deep

learning methods which obtain non-linear latent representation of the data (e.g. scVI[17], scDeepCluster[18], SAUCIE[19], scGNN[20]). Finally, existing methods aimed at extracting interpretable factors from scRNA-seq data (e.g. f-scLVM[21], net-NMFsc[22], Spectra[23]) rely on prior knowledge in the form of known molecular pathways or gene interactions which regulate the desired signal of interest. In contrast, bioIB is applicable in the setting where such prior knowledge is not necessarily available, but the labels with respect to the signal of interest are known. Specifically, bioIB elucidates the relevant gene programs de novo, guided by the cellular labels associated with the signal of interest (e.g., 'Wildtype' and 'Mutant' for the 'Genotype' signal).

In addition to achieving optimal signal-aware clustering of genes via metagenes, bioIB generates a hierarchy of these metagenes, reflecting the inherent data structure relative to the signal of interest. The bioIB hierarchy facilitates the interpretation of metagenes, elucidating their significance in distinguishing between biological labels and illustrating their interrelations with both one another and the underlying cellular populations.

We demonstrate that metagenes generated by bioIB are biologically meaningful, capturing molecular pathways differentially activated between selected cell groups. First, we consider a scRNA-seq atlas of differentiating macrophages. By using either organ-of-origin or developmental stage as signals of interest, we show that bioIB extracts distinct, signal-specific metagene hierarchies and associated biological processes. Next, we demonstrate how bioIB can be used to identify a cellular subpopulation of disease-associated astrocytes in a single nucleus RNA-seq (snRNA-seq) dataset from murine Alzheimer's Disease models. Finally, we showcase that bioIB metagene hierarchy for a dataset of differentiating hematopoietic cell types reflects the developmental hierarchy of the corresponding cellular populations. bioIB is available as an open-source software package, along with documentation and tutorials (https://github.com/nitzanlab/bioIB).

Results

bioIB elucidates signal-specific metagenes and their structure

The bioIB representation is computed for a given dataset and signal of interest, provided as cell labels. The representation is composed of metagenes which are probabilistic aggregation of the genes into clusters, representing the major patterns of gene expression variation underlying the labeled signal.

The input to bioIB includes a count matrix $D \in R^{N\times G}$ of N cells by G genes, and a vector of cell labels related to the signal of interest, $S \in R^{N\times 1}$, where for example, each cell is labeled as sampled from either a healthy or diseased population (Methods; Figure 1A). This input is used to estimate the distributions required for the bioIB algorithm. We thus define three categorical random variables, $C \sim Cat(\{c_1,\ldots,c_N\})$, $X \sim Cat(\{x_1,\ldots,x_G\})$, $Y \sim Cat(\{y_1,\ldots,y_K\})$, respectively representing the N cells, G genes and K cell states of interest. Normalizing the input matrix D by the total number of counts, we obtain a joint probability distribution $p(c,x)$. Next, summing $p(c,x)$ across the cells, we obtain $p(x)$, such that an entry $[p(x)]_i$ represents the marginal probability of sampling the transcript of gene x_i. Using Bayes theorem, we obtain the conditional probability:

$$p(c|x) = \frac{p(c,x)}{p(x)}. \quad [1]$$

Here, an entry $[p(c|x)]_{ij}$ represents the probability of sampling the cell c_i out of all measured cells in D, given that we observed gene x_j .

The provided cellular annotation vector $S \in R^{N\times 1}$ allows us to define the conditional distribution of Y (representing the K cell states of interest) given that we observed a cell in D. By definition $p(y|c)$ is an indicator function, defined by S, namely, for a cell c_i, $p(y|c_i) = 1$ if $S_i = y$ and zero elsewise:

$$[p(y|c)]_{ij} = \mathbb{1}_{S_j = y_i} . \quad [2]$$

At last, we can obtain the conditional distribution of cell states of interest given that we observed a certain gene in D:

$$p(y|x) = \sum_{j=1}^{N} p(y|c_j)p(c_j|x). \quad [3]$$

The conditional probability matrix of cell states given the genes $p(y|x)$ and the gene probability vector $p(x)$ are used as input to the core of the bioIB method, the IB algorithm. The IB yields the optimal probabilistic mapping, $p(\hat{x}|x)$ from the genes' random variable, X, to the categorical random variable representing the metagenes $\hat{X} \sim Cat(\{\hat{x}_1, \ldots, \hat{x}_M\})$, (for $|M| <= |G|$). The mapping is optimal with respect to the tradeoff between compression and information about the signal of interest Y according to a given threshold parameter β (Figure 1C). This is achieved by optimizing for $\hat{X}$ that minimizes the mutual information with the input genes *X,* $I(X, \hat{X})$, while maximizing the mutual information with Y, $I(\hat{X}, Y)$ (Methods; Figure 1C):

$$\hat{X} = argmin_{\hat{X}}(I(X, \hat{X}) - \beta\, I(\hat{X}, Y)). \quad [4]$$

The resulting metagenes are probabilistic clusters of genes capturing the shared expression patterns amongst cell states relative to Y (Figure 1D). The number of metagenes is roughly determined by the threshold parameter β, ranging from the original representation (no compression, $\beta \to \infty$; $\hat{X} = X$) to full compression to a single cluster ($\beta = 0$). A hierarchy of metagenes is obtained by gradually decreasing β through a reverse-annealing process[4] (Methods). The probabilistic output mapping, $p(y \mid \hat{x})$, reflects the amount of information each metagene holds regarding the different labels, whereas the hierarchical structure reveals the interdependence between the metagenes, and the underlying cellular populations they correspond to (Figure 1E).

As an illustrative example, we construct a toy dataset composed of cells belonging to one of two cell types, which act as the signal of interest Y (Supplementary Figure 1A-D). The bioIB hierarchy is revealed by plotting the conditional probabilities $p(y|\hat{x})$ of a particular label given every metagene, across β values that define the compression level (Supplementary Figure 1C-D). The hierarchical structure reflects the interconnections among the metagenes and the specified cell types of interest (Y), while the bifurcation order is dictated by the informativity of the generated metagenes relative to Y.

bioIB can also capture the relationships between related cell types, defined as distinct labels of interest (Y). Given a toy model with four related cell types, bioIB hierarchy reflects the two distinct pairs of linked cell types by two branches. Further splits correspond to higher-resolution separation to different cell types, eventually resulting in cell type-specific metagenes (Supplementary Figure 1E-G).

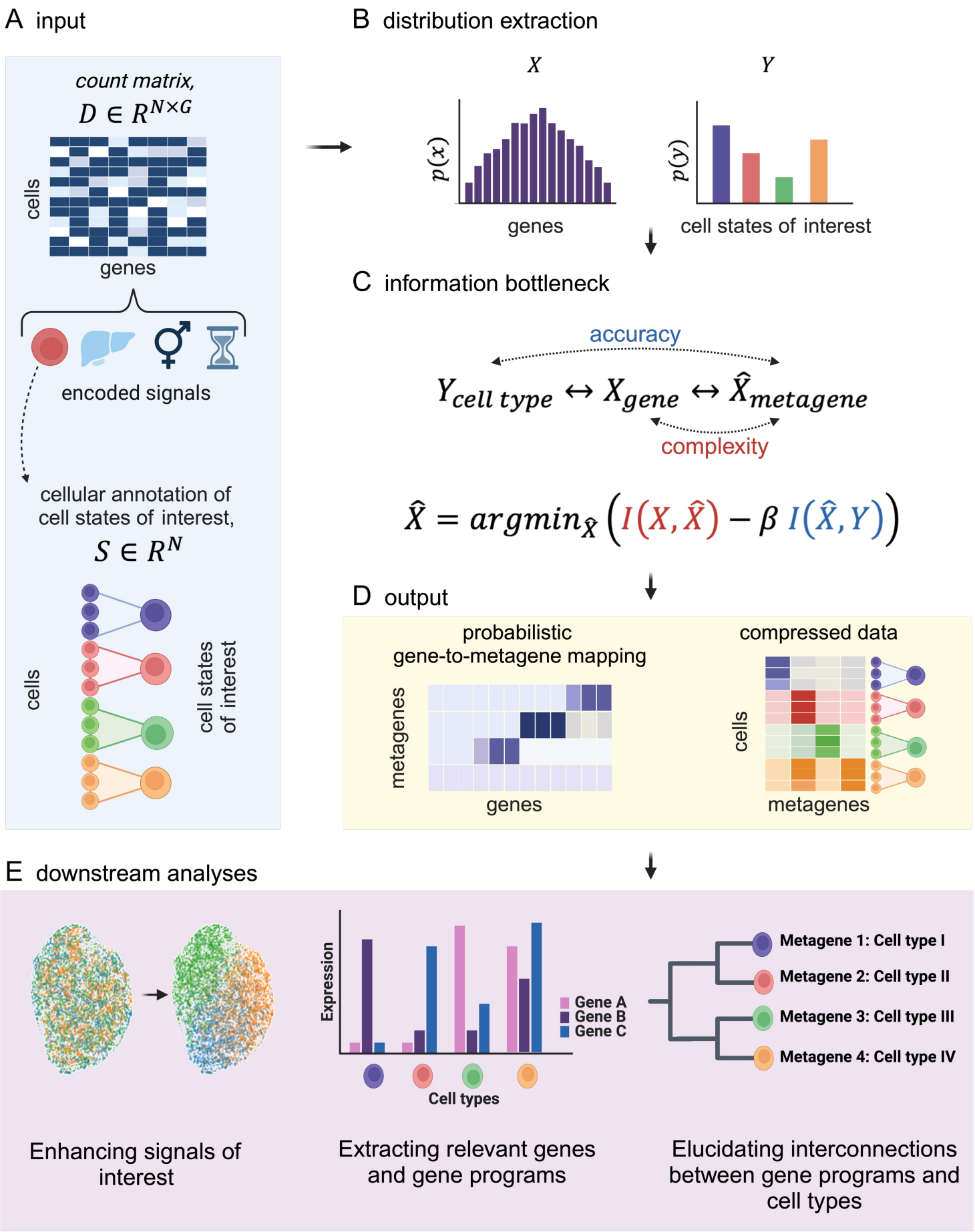
A input
count matrix,
$D \in R^{N \times G}$
cells
genes
encoded signals
cellular annotation of cell states of interest,
$S \in R^{N}$
cells
cell states of interest
B distribution extraction
X
Y
$p(x)$
$p(y)$
genes
cell states of interest
C information bottleneck
accuracy
$Y_{cell\ type} \leftrightarrow X_{gene} \leftrightarrow \hat{X}_{metagene}$
complexity
$\hat{X} = argmin_{\hat{X}} \left(I(X, \hat{X}) - \beta\, I(\hat{X}, Y) \right)$
D output
probabilistic gene-to-metagene mapping
metagenes
genes
compressed data
cells
metagenes
E downstream analyses
Expression
Gene A
Gene B
Gene C
Cell types
Metagene 1: Cell type I
Metagene 2: Cell type II
Metagene 3: Cell type III
Metagene 4: Cell type IV
Enhancing signals of interest
Extracting relevant genes and gene programs
Elucidating interconnections between gene programs and cell types

Figure 1. Elucidating meaningful metagenes underlying the signal of interest using bioIB. A-D) The bioIB pipeline. A) Input; bioIB takes as input a count matrix and a cellular annotation vector, labeling every cell with a state, representing the signal of interest. For example, if the signal of interest is cell type, these labels should annotate every cell with the corresponding cell type. B) Distribution extraction; The provided count matrix and the cellular annotation vector are used to estimate the distributions of the random variables representing the genes (X) and the cell states of interest (Y). C) Information Bottleneck; The probabilities obtained in (B) are used as input for the Information Bottleneck (IB) algorithm, which yields the optimal mapping of genes to metagenes, by optimizing the trade-off between complexity and accuracy. This is achieved by optimizing for $\hat{X}$ that minimizes the mutual information between the input genes (X) and the metagenes ($\hat{X}$), $I(X,\hat{X})$, while maximizing the mutual information between the metagenes and the cell states of interest (Y), $I(\hat{X},Y)$. D) Output; The output of bioIB is a probabilistic mapping between genes and metagenes, scoring all the input genes by their contribution to each metagene. bioIB also provides a cell-to-metagene compressed representation of the input matrix, summarizing the expression of metagenes in the input single cells. E) Possible downstream analyses of the compressed data achieved by bioIB: enhancing the signal of interest (such as cell type or genotype) in the data, extracting informative genes underlying the heterogeneity between cell type labels, elucidating the interconnections between metagenes and the corresponding cell types. Figure is created with BioRender.com.

bioIB extracts distinct molecular signatures in macrophages for developmental stage and organ residence across development

Gene expression data in scRNA-seq experiments contain signatures associated with multiple overlapping biological signals or conditions. How can we identify gene signatures associated with a specific source of heterogeneity in the data? We demonstrate bioIB's approach to this challenge in the context of a scRNA-seq atlas of the developing immune system, which contains cells from 9 organs spanning weeks 4 to 17 after conception[24] (Figure 2A). Applied to this dataset, bioIB identified gene signatures associated with either the organ-of-origin or with developmental stage. For each signal of interest (either organ-of-origin or developmental stage), bioIB clustered the relevant genes into co-expressed metagenes, thus compressing the data, while preserving the information about the selected signal of interest.

We focused on the macrophages subset of cells, as their gene expression varies across organs and throughout the gestation stages[24] (Supplementary Figure 2A). We first applied bioIB with Y set to be the developmental stage, after aggregating cells to 'Early' (8-12 gestational weeks) and `Late' (>14 gestational weeks) labels. The resulting bioIB representation, compressed into four metagenes, exposed the selected signal of developmental stage (Supplementary Tables 1, 2). Qualitatively, cells cluster according to `Early' and `Late' developmental stages in the compressed representation (Figure 2B). Evaluating the corresponding clustering quantitatively we find that the Normalized Mutual Information (NMI) score following bioIB is increased relative to the original data, specifically compared to the developmental stage labels, and not the organ labels (Fold change increase in NMI score: stage labels = 13.8; organ labels = 0.76; Supplementary Table 5). Moreover, bioIB metagenes outperform baseline clustering methods in predicting the gestational week of the cells with an average success rate improvement of 12%, indicating

that bioIB preserves more information about the signal of interest (Figure 2D; Supplementary Figure 3A; Supplementary Table 6; Methods). Applying bioIB to the same data, but now setting the signal of interest Y as the organ-of-origin, resulted in a different compressed representation, informative with respect to the organ-of-origin (Supplementary Tables 3, 4), as the cells in their compressed metagene representation cluster according to their organ-of-origin (Figure 2C). bioIB with Y set to organ improves the NMI scores both for organ clustering and for developmental stage clustering (Fold change increase in NMI score: stage labels = 2.2; organ labels = 1.5; Supplementary Table 5). This is explained by the non-uniform distribution of cells from different organs across developmental stages (Supplementary Figure 2B), such that the information about the cell's organ-of-origin is also highly predictive of its developmental stage. Again, bioIB metagenes outperform baselines in signal label prediction with an average success rate improvement of 12% (Figure 2E; Supplementary Figure 3B; Supplementary Table 7; Methods).

Following the validation of the representation we turn to examine the hierarchical structure and inter-connections of metagenes exposed by the gradual compression of data by bioIB reverse-annealing (Methods).

When Y is set as the gestational stage, as β increases, the first bifurcation represents a metagene that is specific to late gestational stages (Metagene 0; Figure 2F, Supplementary Figure 4A). This metagene is statistically enriched for immune processes based on Gene Ontology analysis[25] (Figure 2G) and contains genes involved in the development of adaptive immune response, such as *TRBV7-2* and *IGL1-40*, that are specifically active in later gestation stages across organs[24] (Supplementary Figure 4B). The next bifurcation produces a more general metagene 1, statistically enriched for multiple biological processes including immune function (Supplementary Figure 4C). Relative to metagene 0, metagene 1 is less informative for the differentiation between early and late stages, and is therefore generated second by bioIB (Supplementary Figure 4D). Metagene 2 contains yolk sac markers (e.g. *TTR*, *CGA*), in accordance with yolk sac being the only organ that is present exclusively at the earliest gestational stages (Supplementary Figure 4B).

Next, we validate that, as desired, metagenes generated with Y being the organ-of-origin reflect organ-specific transcriptomic signatures and that the bioIB hierarchy reflects the developmental connections between the macrophagic populations from different organs (Figure 2H, I, Supplementary Figure 4E). For example, the branch uniting metagenes 0-1, associated with yolk sac, and metagene 2, characterizing liver, is consistent with the reported developmental migration of the macrophages from the yolk sac to the liver[24]. Another example is the branch consisting of metagenes 3 and 4 that mirrors macrophage transition from the liver to the spleen[24]. Together, bioIB extracts distinct, signal-specific hierarchies of metagenes.

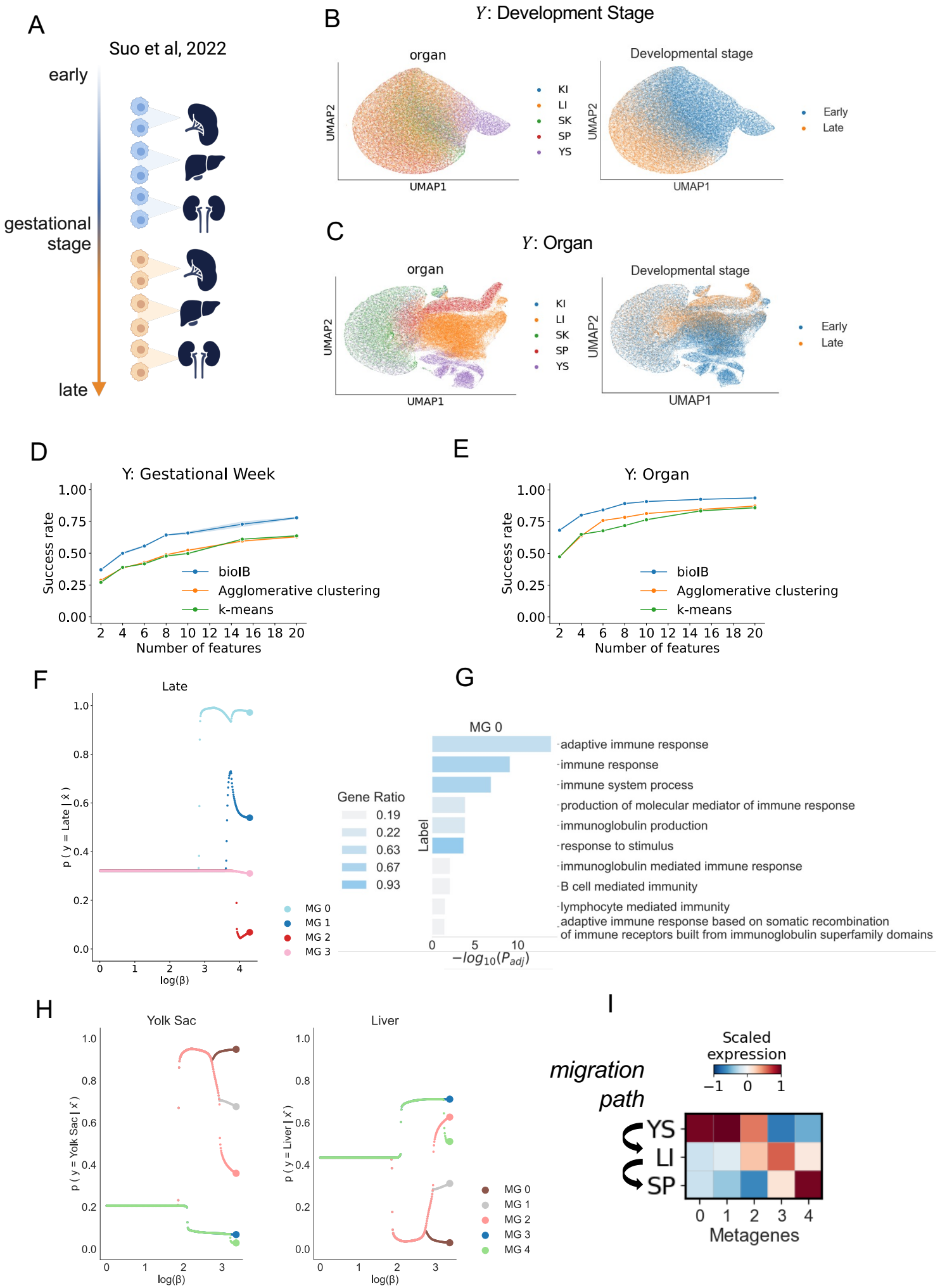
A
Suo et al, 2022
early
gestational stage
late
B
Y: Development Stage
organ
UMAP2
UMAP1
KI
LI
SK
SP
YS
Developmental stage
Early
Late
C
Y: Organ
organ
Developmental stage
D
Y: Gestational Week
Success rate
Number of features
bioIB
Agglomerative clustering
k-means
E
Y: Organ
F
Late
p (y = Late | x̂)
log(β)
MG 0
MG 1
MG 2
MG 3
G
MG 0
Gene Ratio
0.19
0.22
0.63
0.67
0.93
Label
adaptive immune response
immune response
immune system process
production of molecular mediator of immune response
immunoglobulin production
response to stimulus
immunoglobulin mediated immune response
B cell mediated immunity
lymphocyte mediated immunity
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
$-log_{10}(P_{adj})$
H
Yolk Sac
Liver
p (y = Yolk Sac | x̂)
p (y = Liver | x̂)
log(β)
MG 0
MG 1
MG 2
MG 3
MG 4
I
migration path
Scaled expression
YS
LI
SP
Metagenes

Figure 2. bioIB extracts distinct molecular signatures underlying the signals related to developmental stage and organ-of-origin in developing macrophages. A) Schematic representation of the analyzed dataset[24]. The scRNA-seq data of macrophages from 5 distinct organs (kidney, liver, skin, spleen, yolk sac) and 11 gestational weeks (4, 7-12, 14-17), was analyzed using bioIB with Y set either as the developmental stage (Early: < 14 weeks; Late: >= 14 weeks), or as the organ-of-origin, which resulted in two distinct signal-specific compressed representations of the same data. Figure created with Biorender.com. B,C) UMAP representation of the data compressed by bioIB with Y set to developmental stage (B) or organ-of-origin (C), representing single-cell expression of the generated metagenes, colored by the organ (on the left) and by the developmental stage (on the right). D,E) SVM success rates in predicting the gestational week (D) or the organ-of-origin (E) based on bioIB metagenes created with Y set to the respective label; bioIB performance is compared to results based on k-means gene clusters and gene clusters generated by agglomerative clustering based on the log-normalized count matrix X. The x-axis denotes the number of gene clusters or metagenes produced by each method. F,H) bioIB metagene hierarchy produced by reverse annealing with Y set to be developmental stage (F) or organ-of-origin (H), showing the conditional probabilities of the late stage (F) or yolk sac and liver (H) given metagene expression. G) Gene Ontology biological processes significantly enriched among genes representative of metagene 0 shown in (F). In (H), metagene 0 that is mostly expressed in the yolk sac, featuring characteristic organ markers like CGA[26], is merged into one branch with metagenes 1 and 2, that are elevated in the liver and feature apolipoprotein genes involved in liver-associated cholesterol efflux[27]. I) Heatmap showing scaled metagene expression in the original scRNA-seq matrix among organs. YS: yolk sac , LI: liver , SP: spleen.

bioIB metagenes identify Alzheimer's Disease associated astrocytes

A key challenge in scRNA-seq analysis is to identify specific cellular subpopulations affected by a certain condition, such as disease. The standard pipeline, commonly implemented for this task, involves unsupervised clustering of cells, which exposes the downstream analysis to clustering-related bias[28]. BioIB can overcome such limitations and detect disease-associated cells within a heterogeneous cellular population, which we demonstrate in the context of Alzheimer's disease (AD) - associated astrocytes. To do so, we re-analyzed single-nucleus RNA-seq measurements of astrocytes from an AD mouse model and wild-type (WT) mice[29] (Figure 3A).

BioIB analysis with the signal of interest set as the genotype (AD/WT) resulted in a hierarchy of six metagenes (Supplementary tables 9, 10) capturing informative transcriptomic signatures differentiating between AD and WT cells (Figure 3B, Supplementary Figure 5A, Supplementary table 11). The resulting metagenes allowed for better separation of AD and WT cells, compared to the original data (Fold change increase in NMI score = 3.98; Figure 3C).

Furthermore, bioIB metagenes captured a higher-resolution structure within the data; the main branch of metagenes associated with AD phenotype is composed of metagenes 0,1,2, each associated in turn with a distinct subpopulation of AD astrocytes (Figure 3D,E). To interpret their biological identities, we extracted a set of representative genes for each metagene (Methods). Metagene 0, whose representative gene set includes genes involved in morphology regulation (*GFAP, THY1, VIM, B2M, PSEN1*), is enriched for the cellular projection development process, consistent with general astrocyte activation[30]

(Supplementary Figure 5D). Metagenes 1 and 2 represent pathways more tightly associated with the disease: , the representative gene set of metagene 1 is enriched with immune genes[31], such as C1QA and CTSS, and metagene 2 is represented by established markers of AD pathology, *TYROBP* and SERPINA3N[32,33]. Meta-analysis of the AD-associated transcriptome[34] revealed that metagenes 1 and 2 are the only metagenes that are exclusively represented by AD-associated genes (Figure 3F; Methods). Characterization of the WT-related metagene 5 can be found in Supplementary Figure 5E.

While metagene 0 is expressed in the majority of AD astrocytes, metagenes 1 and 2 characterize distinct cellular subpopulations among the AD cells (Figure 3D,E), which we hypothesized to correspond to disease-associated astrocytic signatures. To validate our interpretation, we quantified the expression of bioIB metagenes in six astrocytic clusters defined in the original analysis, which included two homeostatic clusters, two GFAP-high clusters of reactive astrocytes which are not specific to the disease, and two disease-associated clusters[29]. We found that while the bioIB metagene 0 is highly expressed both in disease associated clusters and in reactive GFAP-high clusters, metagenes 1 and 2 are specifically enriched in the disease-associated cluster, most abundant in AD[29] (Figure 3G; Supplementary Figure 5F). The two WT-associated metagenes (4,5) are correspondingly enriched in the homeostatic clusters (Figure 3G). In summary, bioIB allows to directly uncover the cellular subpopulations differentially affected by the disease, avoiding the initial pre-clustering of the cells.

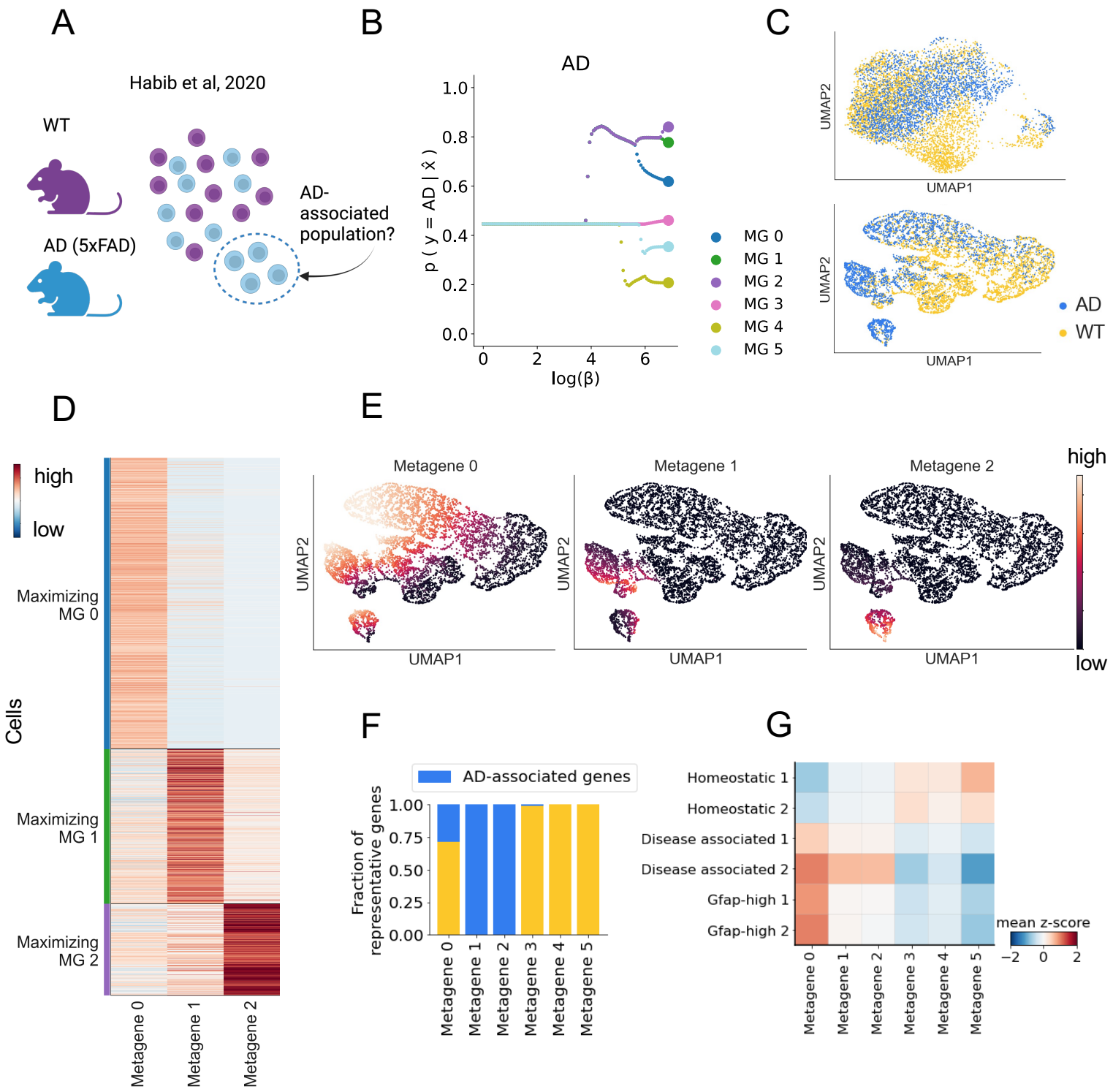
A
Habib et al, 2020
WT
AD (5xFAD)
AD-associated population?
B
AD
p (y = AD | x̂)
log(β)
MG 0
MG 1
MG 2
MG 3
MG 4
MG 5
C
UMAP2
UMAP1
AD
WT
D
high
low
Cells
Maximizing MG 0
Maximizing MG 1
Maximizing MG 2
Metagene 0
Metagene 1
Metagene 2
E
Metagene 0
Metagene 1
Metagene 2
high
low
F
AD-associated genes
Fraction of representative genes
Metagene 0
Metagene 1
Metagene 2
Metagene 3
Metagene 4
Metagene 5
G
Homeostatic 1
Homeostatic 2
Disease associated 1
Disease associated 2
Gfap-high 1
Gfap-high 2
mean z-score

Figure 3. bioIB metagenes identify the AD associated astrocytes. A) Schematic representation of the analyzed dataset[29]. pipeline of the analysis shown in this section. The snRNA-seq data of astrocytes from an AD mouse model was analyzed using bioIB with Y set to genotype, which resulted in identification of a specific subpopulation of disease-associated astrocytes. Figure created with Biorender.com. B) bioIB metagene hierarchy produced given the preprocessed snRNA-seq data, relating to the AD group. The defined metagenes exhibit differential expression patterns between AD and WT, with metagenes 0, 1 and 2 overexpressed in AD cells (Fold change increase in metagenes 0, 1, 2: 1.9, 3.9, 6, respectively), a neutral metagene 3 (Fold change increase in metagene 3 = 0.91), and metagenes 4 and 5 overexpressed in WT cells (Fold change increase in metagenes 4, 5: 0.3, 0.66, respectively; Supplementary table 11). C) UMAP representation of the original data (left) and of the bioIB compressed data (right). D) Heatmap showing scaled expression of metagenes 0,1,2 in individual cells of AD genotype, sorted by maximal normalized metagene expression. E) UMAPs of the bioIB-compressed data, colored by the expression of AD-associated metagenes 0,1,2. F) Fractions of representative genes of metagenes 0-5 that were found to be differentially expressed in at least 7 studies in the meta-analysis of the AD-associated transcriptome[34] (Methods). G) Heatmap of scaled expression values of six bioIB metagenes in six transcriptional clusters of astrocytes, defined in ref[29].

bioIB metagene hierarchy reflects the developmental connections between hematopoietic cell types.

scRNA-seq datasets expose a striking diversity of cell types and states, whose interconnections carry important biological information about cell state identity. For example, the hierarchical differentiation tree of hematopoietic stem and progenitor cells (HSPCs) reveals the phenotype and function of mature hematopoietic cells[35]. BioIB metagene hierarchy can capture the developmental hierarchical structure of cell types, as we demonstrate here for scRNA-seq data of HSPCs differentiation[36] (Figure 4A). BioIB is applied given the cell type signal over a subset of the data containing six major hematopoietic cell types – monocytes, neutrophils, mast cells, basophils, megakaryocytes and erythroid cells. This analysis produced 11 metagenes, where each of the six cell types is uniquely characterized by at least one metagene, maximizing its expression level within that particular cell type (Figure 4B, Supplementary Tables 12,13). In addition, there are metagenes representing a transcriptional program shared by several developmentally linked cell types (Figure 4B,C; Supplementary Figure 6A). For example, metagenes 0 and 2 are specifically expressed in monocytes and neutrophils, respectively, while metagene 1 is activated in both (Figure 4B,C). The bioIB metagenes are biologically informative, uniting genes and processes characteristic of the corresponding cell types (Supplementary Figure 6B,C). Hence, metagene 0, specifically representing monocytes, features monocyte marker genes such as FABP5[36] and WFDC17[37,38] (Figure 4C,D), and is associated with pro-inflammatory macrophage activation, characteristic of monocytes function[39] (Figure 4E). Similarly, metagene 2, specifically characterizing neutrophils, includes markers like ITGB21[36] CAMP, LTF, and ELANE[40] (Figure 4C,D) and is statistically enriched for neutrophil mediated immunity and neutrophil activation (Figure 4E).

The hierarchical representation of the metagenes generated by bioIB induces a hierarchy of cell types that reflects the developmental links between them (Figure 4F,G; Supplementary Figure 6D). In particular, the first bifurcation in the metagene hierarchy generates two metagenes corresponding to the two major branches in the developmental hierarchy[36] (Figure 4A), one which includes Monocytes and Neutrophils, and another which includes Mast cells, Basophils, Megakaryocytes and Erythroid cells (Figure 4F,G). The second bifurcation (Figure 4F,G) splits the latter into two additional specific metagenes, one including Mast cells and Basophils, and another - Megakaryocytes and Erythroid cells (Figure 4F,G). The third bifurcation further splits the metagene corresponding to the Mast-Baso branch to two separate metagenes that are more specific to either Mast cells or Basophils. Similarly, the fourth bifurcation splits the metagene corresponding to the Monocyte-Neutrophil branch to two separate Monocyte and Neutrophil associated metagenes. Finally, the last bifurcations split the metagene corresponding to the Megakaryocyte-Erythroid branch to four metagenes distinguishing between Megakaryocytes and Erythroid cells.

In conclusion, bioIB metagenes characterize distinct biological processes linked to the underlying cellular populations, while the metagene hierarchy unveils the biological relationships interconnecting these populations.

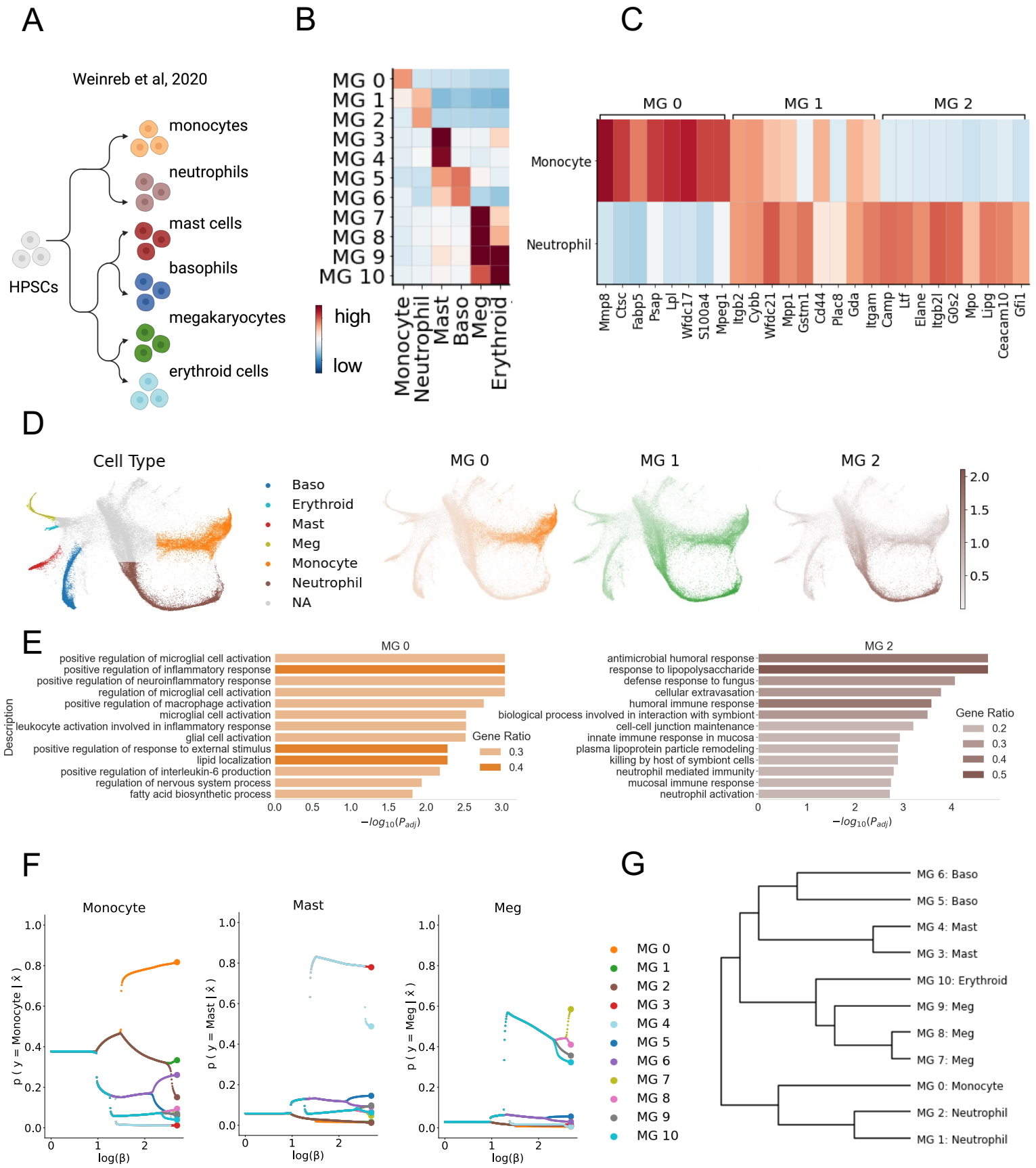
A
Weinreb et al, 2020
HPSCs
monocytes
neutrophils
mast cells
basophils
megakaryocytes
erythroid cells
B
MG 0
MG 1
MG 2
MG 3
MG 4
MG 5
MG 6
MG 7
MG 8
MG 9
MG 10
high
low
Monocyte
Neutrophil
Mast
Baso
Meg
Erythroid
C
MG 0
MG 1
MG 2
Monocyte
Neutrophil
Mmp8
Ctsc
Fabp5
Psap
Lpl
Wfdc17
S100a4
Mpeg1
Itgb2
Cybb
Wfdc21
Mpp1
Gstm1
Cd44
Plac8
Gda
Itgam
Camp
Ltf
Elane
Itgb2l
G0s2
Mpo
Lipg
Ceacam10
Gfi1
D
Cell Type
Baso
Erythroid
Mast
Meg
Monocyte
Neutrophil
NA
MG 0
MG 1
MG 2
2.0
1.5
1.0
0.5
E
MG 0
positive regulation of microglial cell activation
positive regulation of inflammatory response
positive regulation of neuroinflammatory response
regulation of microglial cell activation
positive regulation of macrophage activation
microglial cell activation
leukocyte activation involved in inflammatory response
glial cell activation
positive regulation of response to external stimulus
lipid localization
positive regulation of interleukin-6 production
regulation of nervous system process
fatty acid biosynthetic process
Description
Gene Ratio
0.3
0.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
$-log_{10}(P_{adj})$
MG 2
antimicrobial humoral response
response to lipopolysaccharide
defense response to fungus
cellular extravasation
humoral immune response
biological process involved in interaction with symbiont
cell-cell junction maintenance
innate immune response in mucosa
plasma lipoprotein particle remodeling
killing by host of symbiont cells
neutrophil mediated immunity
mucosal immune response
neutrophil activation
Gene Ratio
0.2
0.3
0.4
0.5
0
1
2
3
4
$-log_{10}(P_{adj})$
F
Monocyte
Mast
Meg
p (y = Monocyte | x̂)
p (y = Mast | x̂)
p (y = Meg | x̂)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
log(β)
G
MG 0
MG 1
MG 2
MG 3
MG 4
MG 5
MG 6
MG 7
MG 8
MG 9
MG 10
MG 6: Baso
MG 5: Baso
MG 4: Mast
MG 3: Mast
MG 10: Erythroid
MG 9: Meg
MG 8: Meg
MG 7: Meg
MG 0: Monocyte
MG 2: Neutrophil
MG 1: Neutrophil

Figure 4. bioIB metagene hierarchy reflects the connections between the developmentally linked hematopoietic cell types. A) Schematic representation of the analyzed dataset[36]. The single-cell data from differentiating blood cells yielded the depicted developmental hierarchy of the hematopoietic cell types. Figure created with Biorender.com. B) Heatmap showing the scaled expression (z-score) of the bioIB metagenes across cell types. C) Heatmap showing scaled expression of the top representative genes of metagenes 0-2 across monocytes and neutrophils. Metagenes 0 and 2 are specifically expressed in monocytes and neutrophils, respectively, while metagene 1 is expressed in both. D) SPRING[41] visualizations of the hematopoietic dataset colored by cell type (left panel) and by the expression of metagenes 0-2 (three panels on the right). **MG – metagene*. E) Gene Ontology enrichment results showing biological process categories significantly enriched in metagene 0 (left) and 2 (right). F) Bifurcation plots of further compression of the 11 metagenes shown in (B) relative to Monocytes, Mast cells and Megakaryocytes. Metagenes characterizing developmentally linked cell types are linked in the metagene hierarchy. For example, metagene 0 representing monocytes diverges from the same branch as metagene 2, representing Neutrophils. Bifurcation plots relative to Neutrophils, Basophils and Erythroid cells are provided in Supplementary Figure 4D. G) Metagene hierarchy inferred from the bioIB reverse annealing output shown in (F) and in Supplementary Figure 4D. The cell type associated with every metagene is the one maximizing the conditional probability of a cell type given this metagene ($max_y p(y|\hat{x})$).

Discussion

We introduced bioIB, a scRNA-seq tailored method for clustering genes with respect to a set of known cellular labels, based on the Information Bottleneck (IB) algorithm. We have shown that bioIB metagenes, which are biologically interpretable, provide a meaningful representation which exposes the representative molecular pathways differentially expressed between cellular populations of interest. Given single-cell data from human differentiating macrophages, with overlapping signals of organ-of-origin and developmental time, bioIB successfully extracted two distinct compressed data representations, each depicting the respective biological processes. Next, we used bioIB to identify a subpopulation of disease-associated astrocytes in single-nucleus data from an Alzheimer's Disease mouse model, providing the genotype as the signal of interest. At last, we have shown that beyond the final metagenes used for analysis, their hierarchical structure, produced by the iterative algorithm used by bioIB, exposes interconnections between metagenes and their respective cell types. We showcased this in the context of differentiating hematopoietic cells, where the bioIB hierarchical structure matched the expected developmental hierarchy of hematopoietic cell types.

By definition, the bioIB output is sensitive to the representation of the differentially labeled cell clusters in the data. The sensitivity is both with respect to the number of genes enriched in each cluster and in terms of the clusters' size . This is expected since the mutual information measure at the core of bioIB is sensitive to these quantities. That is, a cluster containing a small number of cells is expected to be characterized by fewer metagenes compared to a larger cluster. This can be biologically motivated, on the one hand, as the relative abundance of cells within clusters in the data can be associated with their relative

importance, and on the other hand may be a caveat when studying rare populations. For the latter case, a potential solution to this limitation could involve subsampling the data for a uniform distribution of clusters' size.

As with a majority of computational methods, the bioIB output depends on a hyperparameter, β, controlling the level of compression. This is analogous to setting the number of clusters in a clustering algorithm, making this value data-specific. Here, the interpretability of the obtained metagenes allows the user to tune β to obtain the desired number of informative metagenes. The current bioIB formulation is limited in its scalability to data size, as it relies on the exact solution to the IB problem. This can be overcome, as we have done in this study, by focusing the analysis on highly informative genes. A natural extension to bioIB to overcome this limitation more generally is using an existing variational IB solver which relies on neural approximation[42–44].

In future work bioIB can be extended to extract multiple related data representations with respect to several variables of interest, based on the multivariate information bottleneck framework[45]. This paradigm might be particularly useful in analyzing gene expression data, allowing to simultaneously extract multiple encoded signals and analyze the corresponding biological processes. Furthermore, bioIB could be extended to produce signal-specific cell clusters, or metacells, retaining maximal possible information about a target gene subset, such as disease biomarkers.

Here we demonstrated that bioIB can provide efficient characterization of signals of interest encoded in single-cell data, such as cell type, disease state or tissue-of-origin. BioIB can be generalized beyond single-cell gene expression data to additional types of biological data, such as bulk RNA-seq and proteomics data, to expose signal-specific optimally compressed representations. In summary, bioIB is expected to enrich biological data analysis by revealing the hierarchical, signal-specific structure encoded in complex datasets.

Materials and methods

The bioIB algorithm

The bioIB algorithm provides a compressed representation of scRNA-seq data with respect to a signal of interest. To do so it takes as input a cell (N) by gene (G) scRNA-seq measurements matrix, $D \in R^{N \times G}$; following standard practice we suggest providing log-normalized counts as input. Additional input to bioIB is a vector of cell labels related to the signal of interest $S \in R^{N \times 1}$, labeling every cell with one of K possible cell states of interest defined using $Y = [1, \ldots, K]$, such that $Y = \{S\}$. Given this input, the bioIB pipeline is composed of two main steps: (1) obtaining a probabilistic representation of the count matrix, and (2) using this representation as input for the Information Bottleneck (IB) algorithm.

1. Obtaining a probabilistic data representation

We use the input count matrix D and signal of interest S to obtain the relevant probability distributions required for the IB algorithm; the conditional probability matrix of cell states given the genes $p(y|x)$ and the gene probability vector $p(x)$. To convert to probability space we define the random variables of $C \sim Cat(\{c_1, \ldots, c_N\})$, $X \sim Cat(\{x_1, \ldots, x_G\})$, $Y \sim Cat(\{y_1, \ldots, y_K\})$, respectively representing the N cells, G genes and K cell states of interest. The empirical distributions of these are then constructed using the input data (see Equations 1-3).

2. The IB algorithm

The obtained probabilistic representations, $p(y|x) \in R^{K \times G}$ and $p(x) \in R^G$ are the input for the Information Bottleneck (IB) algorithm.

IB[3] is a dimensionality reduction method, designed to extract the information from data X that is relevant for the prediction of another related variable Y, such that the choice of Y determines the relevant components of the signal encoded in X. Mutual information (MI) is used to evaluate both the extent of compression, $I(X, \hat{X})$, and the level of relevant information preserved in the compressed data, through $I(\hat{X}, Y)$. A trade-off parameter β is introduced to control the amount of compression (distortion) allowed. Formally, the IB objective is given by,

$$\hat{X} = argmin_{\hat{X}}(I(X, \hat{X}) - \beta\, I(\hat{X}, Y)).$$

Notably, when $\beta = 0$, all genes are merged into one cluster (full compression), and when $\beta = \infty$, the compressed data is identical to the original full data, so every cluster is associated with one particular gene, $\hat{X} = X$. For every value of β, the algorithm yields the conditional probability matrix of M gene clusters, which we term metagenes, $\hat{x} \in \hat{X}$, given the genes, $x \in X$, $p(\hat{x}|x) \in R^{M \times G}$, representing the optimal mapping of genes to metagenes, and the conditional probability matrix of cell states given the metagenes $p(y|\hat{x}) \in R^{K \times M}$. For the full mathematical description and the associated proofs for the information bottleneck algorithm, see refs[3,4].

There are many ways to solve the IB objective (including neural approximators introduced recently[42–44]). Here we will focus on the Blahut-Arimoto algorithm[46], described below. IB can provide either a series of solutions at different compression levels, using a reverse-annealing process, or a single solution with a flat division of the data points to a predefined number of clusters.

a. Blahut arimoto

$$While\ True:$$

- ➔ $p_{i+1}(\hat{x}|x) = \frac{p_i(\hat{x})}{\sum_{\hat{x}} p_{i+1}(\hat{x}) e^{-\beta D_{KL}[p(y|x)||p(y|\hat{x})]}}\, e^{-\beta D_{KL}[p(y|x)||p(y|\hat{x})]}, \forall\, \hat{x} \in \hat{X}, \forall\, x \in X,$
- ➔ $p_{i+1}(\hat{x}) = \Sigma_x p(x)\, p_{i+1}(\hat{x}|x), \forall\, \hat{x} \in \hat{X}.$
- ➔ $p_{i+1}(y|\hat{x}) = \frac{1}{p_{i+1}(\hat{x})} \Sigma_x\, p_{i+1}(\hat{x}|x)\, p(x, y), \forall\, \hat{x} \in \hat{X}, \forall\, y \in Y.$

$$If\ \forall\, x \in X, \qquad JS_{\frac{1}{2},\frac{1}{2}}[p_{i+1}(\hat{x}|x), p_i(\hat{x}|x)] \leq \varepsilon,$$

$$Break.$$

Here, ε is a threshold parameter used to define convergence based on the difference between previous and current iterations. For a given β, the algorithm converges into a stable solution, providing two output probability matrices that define $\hat{X}$, $p(\hat{x}|x)$ and $p(y|\hat{x})$. $p(\hat{x}|x)$ determines the mapping between the original data points $x \in X$ to data clusters $\hat{x} \in \hat{X}$, whereas $p(y|\hat{x})$ defines the association between the data clusters, $\hat{x} \in \hat{X}$, and the groupings of the signal of interest, $y \in Y$.

b. Reverse-annealing

In the process of reverse-annealing the IB algorithm is initialized with a compressed representation $\hat{X}$ that is identical to the original data X and with a large value of β:

- $p(\hat{x}|x) = I_{|X|}$
- $p(\hat{x}) = p(x)$
- $p(y|\hat{x}) = p(y|x)$
- $\beta_{max} \to \infty$

Next, we run the algorithm iteratively, while reducing β. Upon convergence, we initialize the next iteration with the final $p(\hat{x}|x)$ mapping achieved in the previous step, and with $\beta - \Delta$, for a small step size Δ. Following this procedure we achieve a series of solutions for every value of β: $\forall\beta \in \{\beta_{min}, \beta_{min} + \Delta, \ldots, \beta_{max}\}$. At the end of this process $\beta_{min} \to 0$, corresponding to maximal compression, where $\hat{X}$ consists of a single point, uniting all the original data points in X. Reverse-annealing ultimately yields a hierarchical structure that mirrors several important aspects of the identified clusters, such as their informativity for discrimination between the labels of interest Y, as well as the interconnections among them. It is important to note that β_{max} controls the maximal number of metagenes, namely the number of end-nodes in the hierarchy, and modifying it does not affect the hierarchical structure itself.

c. Clustering

To achieve the division of the data points $x \in X$ to a defined number of clusters M, the IB algorithm is initialized with a random mapping of X to M clusters, generating a binary conditional probability matrix $p(\hat{x}|x) \in R^{M \times G}$. The corresponding $p(\hat{x})$ and $p(y|\hat{x})$ are obtained, using basic probability rules and Bayes Theorem. Since this process introduces a dependence of the output on the initialization, we randomly initialize the algorithm $n = 100$ times and select the mapping that minimizes the objective function (Eq. 1)[4].

Downstream analyses

1. **Identifying representative genes**: The representative genes $x \in X$ for a given metagene $\hat{x}_i \in \hat{X}$ are identified as the ones that maximize $p(x|\hat{x}_i)$. Specifically, for a given metagene, we first order the genes by their conditional probability $p(x|\hat{x}_i)$ ($p(x_1|\hat{x}_i) > p(x_2|\hat{x}_i) >$..). For a given $\tau \in [0,1]$, the set of j representative genes $\{x_1, x_2, \ldots, x_j\}$ is chosen as the minimal set such that:

$$\sum_{k=1}^{j} p(x_k|\hat{x}_i) > \tau.$$

2. **Recovering single-cell metagene expression:** The bioIB output provides the mapping of the original count matrix $D \in R^{N\times G}$ to its compressed representation $\widehat{D} \in R^{N\times M}$. Namely, we obtain the weighted expression of genes, $x \in X$, using the mapping $p(x|\hat{x})$., given by,

$$\widehat{D}_{ij} = \sum_k D_{ik} p\left(x_k|\hat{x}_j\right).$$

   As a result, we obtain a cell (N) by metagene (M) compressed data matrix, $\widehat{D} \in R^{N\times M}$, such that $\widehat{D}_{ij}$ represents the expression level of metagene j in cell i.

3. **Extracting the metagene hierarchy:** The bioIB reverse-annealing output provides a series of conditional probability matrices: $p(x|\hat{x}) \in R^{G\times G}$ and $p(y|\hat{x}) \in R^{K\times G}$ for each β. Since we initialize the reverse-annealing process with $\hat{X} = X$, these matrices include N metagenes, but only M of them are unique. We first identify the most representative gene x of each metagene $\hat{x}_i$, using $p(x|\hat{x})$:

$$x_{\hat{x}_i} = argmax_x p(x|\hat{x}_i)$$

   Next, we extract the metagene hierarchy by identifying the merging points of the most representative genes for each metagene across decreasing β . For example, metagenes $\hat{x}_i$ and $\hat{x}_j$ are considered merged at β_{merge} if $\forall y,\ p(y|x_{\hat{x}_i})_{\beta_{merge}} = p(y|x_{\hat{x}_j})_{\beta_{merge}}$. The identified merging points are recorded using a format of the scipy.cluster.hierarchy.linkage() output linkage matrix and plotted using scipy.cluster.hierarchy.dendrogram(). The code and the documentation for the relevant bioIB functions are provided in the bioIB package at https://github.com/nitzanlab/bioIB.

4. **Linking metagenes to cell types:** Metagenes, $\hat{x} \in \hat{X}$, are linked to cell types, $y \in Y$, using $p(y|\hat{x})$ mapping, given by,

$$y_{\hat{x}_i} = argmax_y p(y|\hat{x}_i).$$

Datasets

Multi-organ atlas of human differentiating macrophages

Data preprocessing

We obtained the dataset of the multi-organ atlas of human differentiating macrophages from ref.[24], available at https://developmental.cellatlas.io/fetal-immune. We downloaded the dataset of myeloid cells and further filtered the data to include only cells of macrophage cell types. Using the gestational week label we assigned the cells into two groups, “Early” and “Late” (Early: < 14 weeks; Late: >= 14 weeks).

In order to avoid bias towards less represented organ groups, we filtered out cells which originated from organs with less than 2800 total cells, resulting in cells originating from five organs: kidney (KI), liver (LI), skin (SK), spleen (SP) and yolk sac (YS). Following basic preprocessing for low-quality cells using scanpy's[47] `sc.pp.filter_cells(min_genes=200)`, the data used for bioIB analysis included 108197 cells. We further reduced the data to 500 highly variable genes using scanpy's[47] `sc.pp.highly_variable_genes()` with the default parameters.

Method application

We applied bioIB to the obtained dataset twice, (1) setting Y as the development stage ($Y = [Early, Late]$), and (2) setting Y as the organ-of-origin ($Y = [KI, LI, SK, SP, YS]$). In both analyses we initialized the reverse-annealing process with $\beta_{max} = 20$.

Benchmarking

To evaluate the predictive performance of bioIB metagenes, we trained linear SVMs on the compressed data generated by bioIB and compared the success rates in predicting the cellular labels of gestational week and organ with the ones achieved by two settings of k-means gene clusters, computed over (1) the original log-normalized count matrix *X,* and (2) the probability matrix *p(y|x),* that contains the information about the labels and represents the original input to bioIB.

Astrocytes from a murine model of Alzheimer's Disease (AD)

Data preprocessing

We obtained single-nucleus RNA-seq measurements from astrocytes from AD mouse model and wild-type (WT) mice from ref.[29], available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143758. Following normalization and log-transformation, we performed leiden clustering using Scanpy's `sc.tl.leiden()' function with default parameters. Following this, we retained the cell clusters with enriched expression of the astrocytic markers *Gfap* and *Slc1a3*, resulting in n=7036 cells. As a last step we extracted highly informative genes with respect to the signal of interest, or disease state, encoded by the provided genotype annotation $Y = [AD, WT]$. This was done by retaining the 1000 genes with the highest information gain (IG) values, where the IG is defined using the mutual information between the gene expression probability $p(x)$ and the genotype probability $p(y)$,

$$IG\ (x) = p(x)\ D_{KL}\ (p(y|x)\ ||\ p(y)\ ).$$

Method application

We applied bioIB with Y set as the mouse genotype: $Y = [AD, WT]$. The reverse-annealing process was initialized with $\beta_{max} = 150$.

Constructing a list of AD-related genes

AD-associated genes were defined as differentially expressed genes in at least 7 of the 15 AD-APP mouse model studies as part of the AD meta-analysis resource, which has summarized and compared the differential expression results from a wide range of AD transcriptomic studies[34].

Hematopoiesis dataset

Data preprocessing

We obtained the dataset of the differentiating hematopoietic cell types collected by ref.[36] and processed by ref.[35]. Data was downloaded using the Cospar package (https://cospar.readthedocs.io/en/latest/index.html) using the function `cs.datasets.hematopoiesis()`). We filtered out the undifferentiated cells, as well as the differentiated cell types with less than 300 total cells, resulting in a data subset of 27387 cells. We further reduced the data to the highly variable genes using scanpy's[47] `sc.pp.highly_variable_genes()` with the default parameters, resulting in 1803 genes.

Method application

We calculated the IG values (as above) for the highly variable genes and used as input for bioIB the 300 genes with the highest IG values.

Data availability

The datasets analyzed in the current study are available at:

- Immune macrophage atlas: https://developmental.cellatlas.io/fetal-immune
- Alzheimer's Disease astrocytes: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143758
- Hematopoiesis: https://cospar.readthedocs.io/en/latest/index.html

Code availability

Software is available at https://github.com/nitzanlab/bioIB.

Acknowledgements

We would like to thank the late Professor Naftali Tishby for initiating this project and his guidance which made this work possible. We would also like to express our gratitude to Professor Eli Nelken and Hadar Levi Aharoni for fruitful discussions. We acknowledge all members of the Nitzan lab for general feedback. This work was supported by the Azrieli, Kaete-Klausner and TEVA PhD fellowships (S.D.), a scholarship for outstanding doctoral students in data-science by the Israeli Council for Higher Education and the Clore Scholarship for PhD students (Z.P.), an Alon Fellowship, the Israel Science Foundation (Grant no. 1079/21),

and the European Union (ERC, DecodeSC, 101040660) (M.N.). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council.

References

1. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610–620 (2015).
2. Berger, T. Rate-Distortion Theory. in *Wiley Encyclopedia of Telecommunications* (ed. Proakis, J. G.) (Wiley, 2003). doi:10.1002/0471219282.eot142.
3. Tishby, N., Pereira, C. & Bialek, W. The Information Bottleneck Method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* **49**, (2001).
4. Slonim, N. The Information Bottleneck: Theory and Applications. *Ph.D Thesis* (2002).
5. Slonim, N. & Tishby, N. Document clustering using word clusters via the information bottleneck method. in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* 208–215 (ACM, Athens Greece, 2000). doi:10.1145/345508.345578.
6. Gordon, Greenspan, & Goldberger. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. in *Proceedings Ninth IEEE International Conference on Computer Vision* 370–377 vol.1 (IEEE, Nice, France, 2003). doi:10.1109/ICCV.2003.1238368.
7. Tan, A. K., Tegmark, M. & Chuang, I. L. Pareto-Optimal Clustering with the Primal Deterministic Information Bottleneck. *Entropy* **24**, 771 (2022).
8. Zaslavsky, N., Kemp, C., Regier, T. & Tishby, N. Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 7937–7942 (2018).

9. Schneidman, E., Slonim, N., Tishby, N. & Bialek, W. Analyzing Neural Codes Using the Information Bottleneck Method. (2001).

10. Jin, B. & Lu, X. Identifying informative subsets of the Gene Ontology with information bottleneck methods. *Bioinformatics* **26**, 2445–2451 (2010).

11. Bauer, M. & Bialek, W. Information Bottleneck in Molecular Sensing. *PRX Life* **1**, 023005 (2023).

12. Bauer, M., Petkova, M. D., Gregor, T., Wieschaus, E. F. & Bialek, W. Trading bits in the readout from a genetic network. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2109011118 (2021).

13. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine* **59**, 114–122 (2018).

14. Rostom, R., Svensson, V., Teichmann, S. A. & Kar, G. Computational approaches for interpreting sc RNA -seq data. *FEBS Letters* **591**, 2213–2225 (2017).

15. Wang, W. *et al.* Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Research* **49**, e54–e54 (2021).

16. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4164–4169 (2004).

17. 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).

18. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat Mach Intell* **1**, 191–198 (2019).

19. Amodio, M. *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat Methods* **16**, 1139–1145 (2019).

20. Wang, J. *et al.* scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat Commun* **12**, 1882 (2021).

21. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* **18**, 212 (2017).

22. Elyanow, R., Dumitrascu, B., Engelhardt, B. E. & Raphael, B. J. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res.* **30**, 195–204 (2020).

23. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01940-3.

24. Suo, C. *et al.* Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).

25. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

26. Gulbis, B., Jauniaux, E., Cotton, F. & Stordeur, P. Protein and enzyme patterns in the fluid cavities of the first trimester gestational sac: relevance to the absorptive role of secondary yolk sac. *Molecular human reproduction* **4**, 857–862 (1998).

27. Fielding, C. J. & Fielding, P. E. Cellular cholesterol efflux. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1533**, 175–189 (2001).

28. Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2100293118 (2021).

29. Habib, N. *et al.* Disease-associated astrocytes in Alzheimer's disease and aging. *Nat Neurosci* **23**, 701–706 (2020).

30. Pekny, M., Wilhelmsson, U. & Pekna, M. The dual role of astrocyte activation and reactive gliosis. *Neuroscience Letters* **565**, 30–38 (2014).

31. Sekar, S. *et al.* Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiology of Aging* **36**, 583–591

(2015).

32. Zhao, N. *et al.* Alzheimer's Risk Factors Age, APOE Genotype, and Sex Drive Distinct Molecular Pathways. *Neuron* **106**, 727-742.e6 (2020).

33. Park, S. *et al.* Gene expression profiling of aging in multiple mouse strains: identification of aging biomarkers and impact of dietary antioxidants. *Aging Cell* **8**, 484–495 (2009).

34. Wan, Y.-W. *et al.* Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. *Cell Reports* **32**, 107908 (2020).

35. Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat Biotechnol* **40**, 1066–1074 (2022).

36. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).

37. D'Souza, S. S. *et al.* Type I Interferon signaling controls the accumulation and transcriptomes of monocytes in the aged lung. *Aging Cell* **20**, e13470 (2021).

38. Kasmani, M. Y. *et al.* A spatial sequencing atlas of age-induced changes in the lung during influenza infection. *Nat Commun* **14**, 6597 (2023).

39. Italiani, P. & Boraschi, D. From Monocytes to M1/M2 Macrophages: Phenotypical vs. Functional Differentiation. *Front. Immunol.* **5**, (2014).

40. Xie, X. *et al.* Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. *Nat Immunol* **21**, 1119–1133 (2020).

41. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).

42. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep Variational Information Bottleneck. in *International Conference on Learning Representations* (2017).

43. Belghazi, M. I. *et al.* Mutual Information Neural Estimation. in *Proceedings of the 35th International Conference on Machine Learning* (eds. Dy, J. & Krause, A.) vol. 80 531–540 (PMLR, 2018).

44. Wieczorek, A. & Roth, V. On the Difference between the Information Bottleneck and the Deep Information Bottleneck. *Entropy* **22**, 131 (2020).

45. Friedman, N., Mosenzon, O., Slonim, N. & Tishby, N. Multivariate Information Bottleneck. (2013) doi:10.48550/ARXIV.1301.2270.

46. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory* **18**, 460–473 (1972).

47. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).