Disentangling the different attributes shaping the development of the human immune system

Undergraduate Student: Bar Melinarskiy[1]

Instructors: Zoe Piran[1] and Dr. Mor Nitzan[1,2,3]

August 2023

Abstract

The intricate development of the human immune system is a critical biological process essential for safeguarding against pathogens and maintaining overall health. Progress in the understanding of this system can be attributed in part to the advancements in sequencing techniques like scRNA-seq. As this technology enables gathering data on gene expression within specific cell types and subpopulations. In this study, we showcase the abilities of biolord, a deep generative framework for learning disentangled representations in single-cell data in the context of the development of the immune system. This framework leverages recent advancements in disentanglement from the computer vision field by extending deep learning frameworks that rely on latent optimization. In particular, we relied extensively on biolord's data manipulation and predictive capabilities. When trained on a single-cell dataset, this tool generates a disentangled latent representation of known and unknown attributes. Then, through its embedded generator, we can manipulate known attributes of specific cells and predict the resultant impact on their gene expression.

Using biolord on human embryo B-cell data, distinct gene expression patterns linked to cell types and organs were unveiled. Focusing on B-cell migration, we shifted Pre-pro B cells from the yolk sac to other immune-related organs, revealing gene expression shifts tied to organ identity and highlighting the spleen's pivotal role in B cell maturation. Similarly, shifting B cell types' revealed gene profiles related to stress responses, metabolic regulation, and immune function. This study underscores computational models' potential, like biolord, in shedding light on the pivotal biological processes in a single-cell resolution.

1 The Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel

2 Racah Institute of Physics, The Hebrew University, Jerusalem, Israel

3 Faculty of Medicine, The Hebrew University, Jerusalem, Israel

Introduction

The developing immune system

The human immune system's development is a vast, intricate biological process, vital for defense against pathogens and overall health maintenance. Figure 1A illustrates a complex sequence of steps, orchestrated by diverse organs and cell types to establish a robust defense mechanism. Immune cells stem from extra-embryonic yolk sac progenitors and hematopoietic stem cells from aorta-gonad-mesonephros. The liver and bone marrow become primary hematopoietic sites, seeding immune cells in lymphoid and peripheral nonlymphoid organs, laying the groundwork for immune function [1].

The thymus, bone marrow, lymph nodes, and spleen are pivotal in immune system development, driving T cells, B cells, Macrophages, Natural Killer (NK) cells, and dendritic cells maturation. Notably, B cells play a central role in generating antibodies and ensuring enduring immunity. Understanding immune system intricacies empowers researchers and medical experts to enhance responses, create vaccines, and tailor therapies for immune-related disorders.

B cell development

B cell development, crucial for the immune system, undergoes a series of key steps to ensure functionality. Originating in the fetal liver, it progresses to the bone marrow, where hematopoietic stem cells differentiate into early B cell progenitors. This leads to maturation marked by immunoglobulin gene rearrangement, enabling B cells to generate diverse B cell receptors (BCRs) via V(D)J recombination. Successful BCR gene rearrangement results in immature B cells, which then migrate to secondary lymphoid organs like the spleen and lymph nodes [2–4].

In secondary lymphoid organs, additional selection occurs. Immature B cells interact with antigens presented by dendritic cells. Positive selection boosts cells whose BCRs bind antigens effectively, ensuring broad pathogen recognition. Conversely, negative selection eliminates cells with self-reactive or weakly binding BCRs via apoptosis [4–6]. Suo et al.'s recent

work [1] unveils broader immune cell development beyond conventional sites, revealing B lymphopoiesis across peripheral organs (Gut, Skin, Kidney, etc.) as in Figure 1B.

In summary, immune system development involves intricate coordination among organs and cell types. B cell maturation is pivotal, ensuring diverse BCRs for pathogen recognition. Understanding B cell development aids immunity comprehension and offers avenues for innovative immunotherapies and vaccines against diseases.

Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is a revolutionary method, transforming gene expression analysis to the individual cell level. Unlike bulk RNA sequencing, scRNA-seq captures distinct gene expression patterns within cell types and subpopulations. It's essential for understanding cellular diversity, development, and diseases, complementing other sequencing methods [7].

Disentanglement and latent representation learning

Latent representation learning refers to the process of transforming raw data (e.g. an image or single-cell data) into a compact and meaningful representation, often in a lower-dimensional space. Given a set of input data points X, the goal is to learn a function f that maps X to a set of latent representations Z_X: $Z_X = f(X)$ [8,9]. These latent representations are designed to capture formative features of the data, making it easier to perform various downstream tasks like clustering, classification, or generation. The process of learning these latent spaces involves extracting relevant information from the data while reducing noise and irrelevant variations. Common techniques for achieving latent representation learning include autoencoders, and generative adversarial networks (GANs) [8]. In the realm of biology and translational medicine, this concept finds application in diverse contexts, including genetic data analysis, biomarker identification, and personalized medicine strategies development [8].

When people observe an object, they naturally try to grasp its various attributes like shape, size, and color, using prior knowledge. However, current deep learning models often directly learn how to represent the object to fit the data patterns and discrimination criteria [10]. Unfortunately, this approach often misses the ability to capture the underlying

characteristics hidden in these representations, which are crucial for human-like generalization. To this end, a fairly new approach to representation learning comes into play: Disentangled Representation Learning [11,12]. As such, disentanglement and latent representation learning are closely interconnected tasks. Disentanglement involves separating the different factors of variation within the data, a task that surely can leverage latent representation learning [8,11,12]. Originating from image processing, the disentanglement task in image analysis consists of recovering the compositions of permanent (e.g., class identity) and transitory (e.g., pose) attributes [13]. For example, as shown in Figure 1C, presents a simulated image toy example taken from DeepMind's latest neural network MONet [12]. One can clearly see that the image consists of distinct shaping factors, such as a yellow background, green ground, and various shapes within it. These factors can be encoded individually using an encoder, which results in a learned latent space where each attribute is represented separately. Once we obtain the learned representation, we can manipulate individual or multiple attributes within this space. Finally, by utilizing a generator, we can then decode this matrix back to the image space and analyze the effects of the changes on the image. Through this process, we can enhance our understanding of how each shaping factor contributes to the overall topology of the image.

biolord (biological representation disentanglement)

In this work, we used biolord, a deep generative framework for learning disentangled representations in single-cell data, recently presented by Piran et al [12]. This framework leverages recent advancements in disentanglement from the computer vision field by extending deep learning frameworks that rely on latent optimization [12,14,15]. As a semi-supervised learning model, it requires the training set to consist of single-cell measurements, each with partial supervision over a limited set of known attributes [12].

For example, in our case the known attributes are the cell type labels, the source organ, and the donor's age and sex; attributes may be categorical (discrete; e.g. cell type) or ordered (continuous; e.g. age). As such, we obtain a disentangled representation for all (known and unknown) attributes in the data [9]. On top of these, biolord learns a generator, which maps the representations of the known and unknown attributes into observable single-cell data. It can, in turn, use the disentangled latent space to predict single-cell measurements for different cell

states across variations in internal or external conditions [12], this ability is the basis for this project. Here we focus on disentangling the attributes in the development of the human immune system, specifically focusing on the maturation and migration processes of B-cells and utilize biolord's predictions to explore gene expression shifts associated with organ identity as well as its cell type. For a more in-depth explanation of this tool and its training process in this project, refer to the methods section.

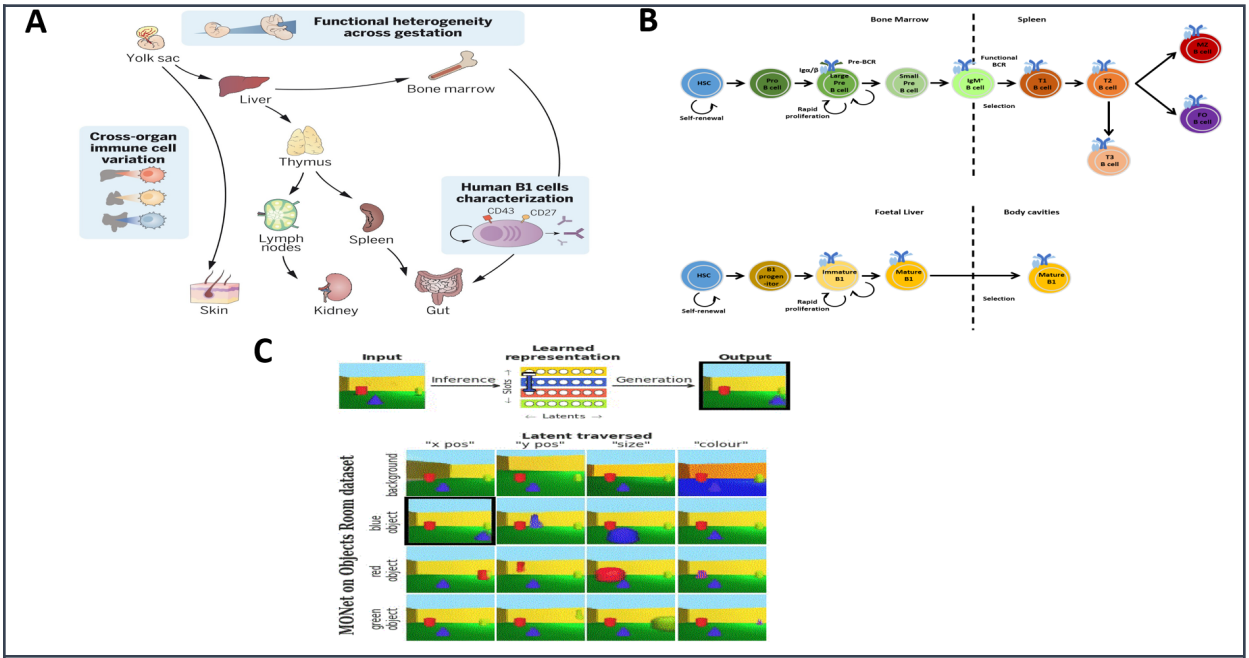


Figure 1. (A) A map, sourced from Suo et al [2], charts immune cell development across prenatal hematopoietic, lymphoid, and peripheral organs. It combines scRNA-seq, scVDJ-seq, and spatial transcriptomics data, integrating dissociated cells from the yolk sac, prenatal spleen, and skin, along with publicly available cell atlases from six other organs. The study spans 25 human embryos at weeks 4 to 17 after conception. (B) B cells maturation stages, generating antigen-specific immunoglobulins (Ig) to combat pathogens. B cell development initiates in the fetal liver, continues in the bone marrow, and migrates to the spleen as IgM+ cells. (C) A toy example of the disentanglement of an image was taken from DeepMind's latest neural network MONet [16]. One can see how the interface perfectly separates each object from the background. Such skills are fully unlocked during unsupervised learning. Once the disentangled latent representation is obtained we can manipulate each attribute encoded and inspect how the change has affected the original image.

Results

The data used for this project was taken from the work of Suo et al [1]. In their study, scRNA-seq was applied to cells from 25 human embryos at 7 to 14 weeks postconception (pcw)

across nine organs, see Figure 1A. Similarly to the toy example (Figure 1C), we aim to gain a representation of the disentangled immune cell attributes, known and unknown. Instead of an image, our input is the cells' gene expression and its known attributes: type, organ identity, and the age and sex of the donor. Just like the image encapsulated various attributes, the single cell's gene expression encodes attributes like cell type, and tissue (Figures 2A and B), and identifying the contribution of each attribute to the gene expression profile is indeed a challenging task.

Yet, uncovering this disentangled representation could reveal gene trends and their related biological processes. As mentioned above, once this representation is obtained, biolord enables us to shift cells' known attributes, in our case organ and cell type, and predict their gene expression in this new state [12]. Thus, we could simulate processes of cells' migration and maturation happening in our developing immune system and uncover related gene trends. A more detailed explanation of how biolord was trained in this project can be found in the methods section.

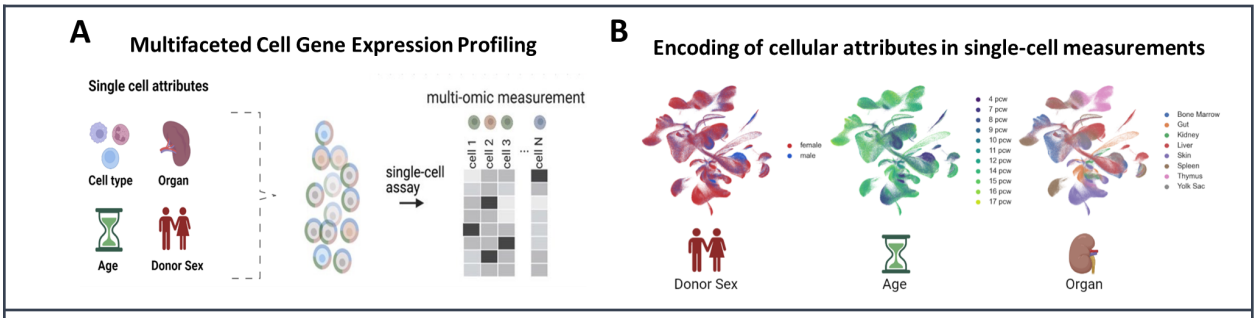


Figure 2. (A) A cell's gene expression profile simultaneously encodes information about multiple attributes, such as cell type, tissue of origin, and the donor's age. (B) UMAPs of gene expression of 64KB-Cells. Data was taken from the study of Suo et al. [1]. Color-coded by the following cell attribute from left to right: donor Sex, age, and organ.

Unveiling the Distinct Gene Expression Profile of B-Cells in the Spleen

As aforementioned, one of our main goals is to identify the gene trends related to the cells' organ identity. To achieve this, biolord [12] was trained on a dataset of B-Cells sourced from the research of Suo et al [1]. Following data pre-processing, 64K cells remained for training biolord (Code available at biolord_immune_bcells). The biolord framework offers versatile

control over the training process. For simplicity, we treated known cell attributes – organ, cell type, and age of donor – as discrete values. Upon completing the training phase, the trained model revealed a decomposed latent space, featuring informative embeddings for known attributes and an embedding for remaining unknown attributes. Then the generative module can harness this decomposed latent space to predict single-cell measurements across diverse cell states [12]. Thus, utilizing a trained biolord model, allows us to make counterfactual predictions and gain the gene expression of the changed cells [12]. This approach facilitates the exploration of gene trends relative to specific cell attributes, specifically here organ-related gene trends.

With our trained model ready, we delved into the decomposed latent space (Figure 3A). This exploration unveiled unique clusters, with some B-Cells clustering based on their source organs, such as bone marrow, spleen, and skin cells. Similarly, cell types like Plasma B, Immature B, and Small pre-B exhibited clustering patterns as well. However, some cells did not exhibit specific patterns in the formation of their clusters. Focusing on the spleen as a representative case, we used the trained model to relocate Pre-pro B cells from the yolk sac to the spleen. This showcases biolord's ability to predict gene expressions for a control cell's population shifted from their original sampled organ to a new target organ, a feat challenging to achieve experimentally due to technical limitations.

The process of inferring changes, at the individual cell level, induced by shifts in a cell's source organ is depicted in Figure 3B. Upon acquiring the predicted gene expressions of cells moved from the yolk sac to the spleen, we proceeded to perform statistical analysis. In this case, we employed a dependent t-test for paired samples. This test aims to assess the null hypothesis that the altered predictions and the original gene expressions share identical average expected values. We repeated this evaluation for each gene among the 4192 genes that persisted in the pre-processing steps. Subsequently, the outcomes of these tests serve as input for gene set enrichment analysis (GSEA) [17]. This analysis revealed the activation of gene pathways linked to proliferation and the regulation of Lymphocyte cells, and also suppression of biosynthetic-related pathways, as illustrated in Figure 4A. These findings are in accordance with previous reports [18–20].

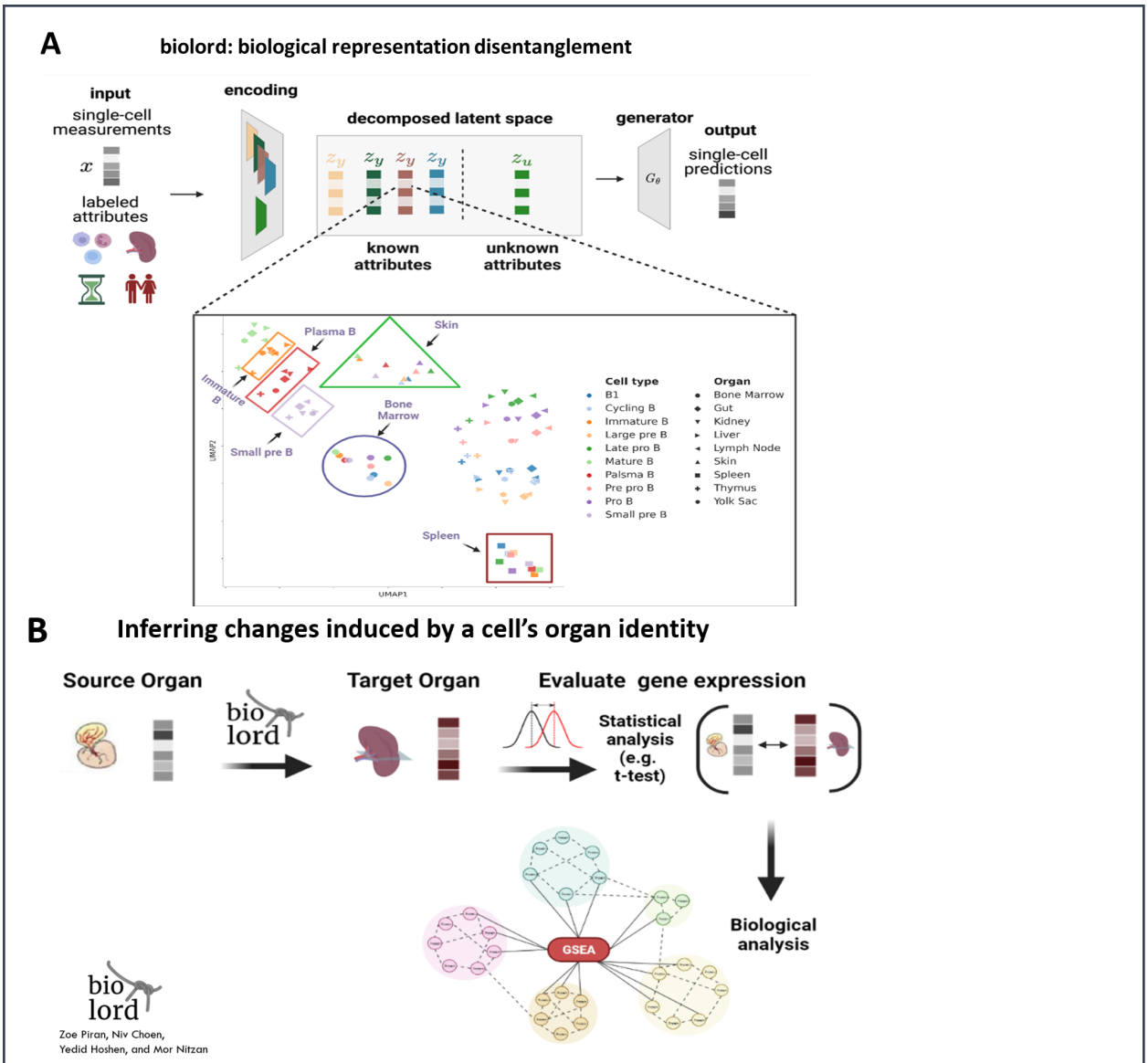


Figure 3. (A) A UMAP [21] showcasing the embedded learned cell attributes. Each individual dot corresponds to a distinct cell type within a specific organ. The color of a dot is linked to its corresponding cell type, while its shape is associated with the unique organ identity. To highlight the formed clusters, we've introduced distinctive shapes like the green triangle onto the plot. (B) The pipeline we used for utilizing biolord for downstream analysis tasks. We can study the changes in gene expression that correspond to a manipulation of a cell's attribute. In the presented case, we illustrate predicting the gene expression of B-Cells obtained from the yolk sac as though they were sampled from the spleen. With both the initial and modified gene expression profiles in our hands, we can perform statistical examination (e.g. t-test) to understand which genes were upregulated and which were downregulated. Then, in order to gain more generalized conclusions on the affected gene pathways and related biological processes we cross-reference with known expression profiles databases using GSEA.

Immunotherapy potential of B cells' targeted strategies

To get a deeper understanding of the gene expression profile of B cells residing within the spleen, we zoomed in on a specific set of genes (listed in Figure 4B) which are recognized as markers of early-specific and late-specific NK neighborhoods, taken from Figure 3B in Suo et al.'s work [1]. Our assumption is that these genes, known for distinguishing between organs in the context of NK cells [1], will offer insight into organ-specific biological processes related to B-Cells. This set includes 40 genes categorized into four groups: Control random genes (10), immune function-related genes (5), inflammatory response-related genes (4), and TNF signaling genes (21). While the random control group displayed minimal variation across organs, as expected, the remaining three gene groups demonstrated noticeable variations. Notably, many genes from TNF signaling and Immune function categories exhibited elevated expression predominantly within the spleen. Conversely, a majority of Inflammatory response-related genes showed relatively diminished expression levels, particularly in the spleen.

The downregulation of inflammatory response-related genes like CD70, CXCR5, and RGS16 in spleen-residing B cells may be attributed to complex regulatory mechanisms that maintain immune homeostasis and prevent excessive immune activation [22–24]. For instance, the inhibition of CD70, a gene crucial for B-cell activation and T-cell co-stimulation, might relate to the mitigation of uncontrolled B-cell activation and subsequent immune-mediated tissue damage [24]. Similarly, the reduced expression of CXCR5, integral for B-cell migration to germinal centers, possibly indicates moderation of B-cell trafficking to prevent excessive immune reactions and autoimmunity [23]. Finally, suppression of RGS16, a G-protein signaling regulator, may prevent B-cell hyperactivation and inappropriate immune responses [22]. These findings align with established regulatory mechanisms that fine-tune the immune system, promoting balance rather than excessive activation or suppression [25,26].

Moving on, we examined the KLF, FOS, and JUN genes within the TNF signaling group revealing a distinct pattern. These genes exhibit elevated expression in the spleen while being suppressed in other organs, consistent with prior studies [27].

The Fos gene family encodes leucine zipper proteins that dimerize with JUN family proteins, forming the AP-1 transcription factor complex. AP-1 regulates cell proliferation,

differentiation, and transformation, and has been linked to apoptotic cell death in B-cells in the spleen [28,29]. In accordance, genes from the KLF family, associated with apoptosis, also show spleen-specific overexpression. While remembering the spleen's role as pivotal in the transition from immature to mature B cells, we note that a systematically programmed cell death among B-cells is a natural occurrence during a state of homeostasis [29] and indeed our data derives from healthy human embryos.

These insights demonstrate computational tools' capability to reveal hidden biological processes. When we gain a deeper understanding we can access better treatment when things go wrong. Such is the case with the gained understanding of the role of AP-1 in B-cell selection which is leveraged in autoimmune diseases using Dihydroartemisinin (DHA), a potent antimalarial drug. Li et al. found DHA boosts Treg while suppressing B-cells in germinal centers. Attributed to the upregulation of c-Fos expression by DHA and the enhancement of its interaction with target genes in both Treg and circulating plasma cells with bilateral cell fates [28]. Refer to Figure 4C for a schematic of the docking interactions between human c-Fos-c-Jun with DNA complex [28].

Gene Expression Profiling of Early Human B Cell Development Stages

Once more, we put biolord's data manipulation and prediction abilities to work, but this time our focus was on uncovering gene trends related to the cell's type. We first simulated the development of Immature B-cells, sampled from the liver at 16 pcw to Mature B-cells. All the while, we kept all other known attributes unchanged in order to be able to directly attribute changes in the predicted gene expression of the counterfactual Mature B-cells to the transition in the cells' type.

Similarly to the analysis done on splenic cells, we performed dependent t-tests for paired samples, comparing mean expression levels between source and target cell types. The subsequent GO-GSEA analysis revealed activated gene pathways intricately tied to the cell type shift, as depicted in Figure 5A. Notably, these activations aligned with intracellular transport, cellular stress response, and metabolic regulation. These results serve as compelling support, as we know that the developmental journey of B-cells exposes them to diverse environments marked by fluctuating nutrient and oxygen concentrations, pH levels, and temperature. All

these dynamic conditions influence B cell signaling and metabolism. Evidently, the readiness of B cells to counteract such stressors is pivotal for their vitality and optimal functionality [25,26].

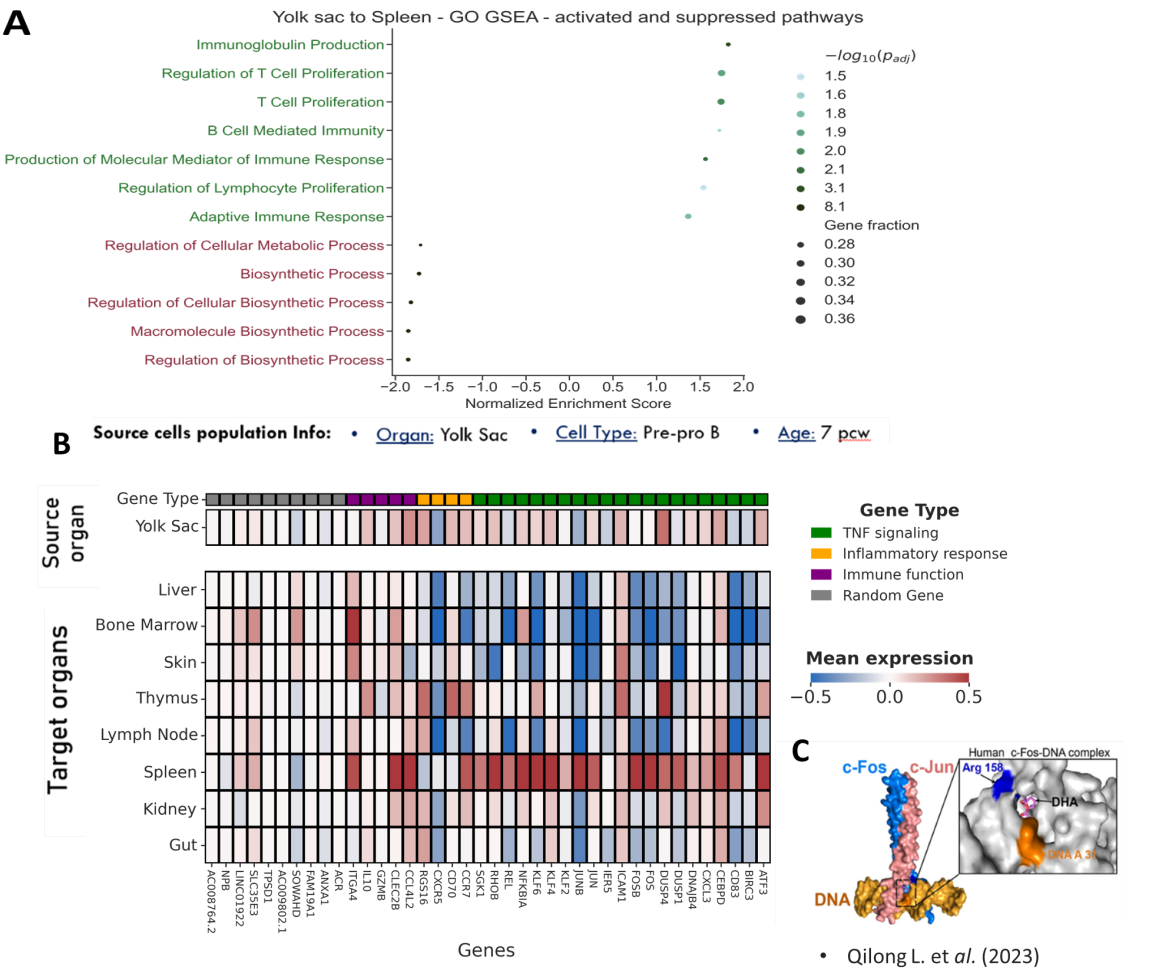
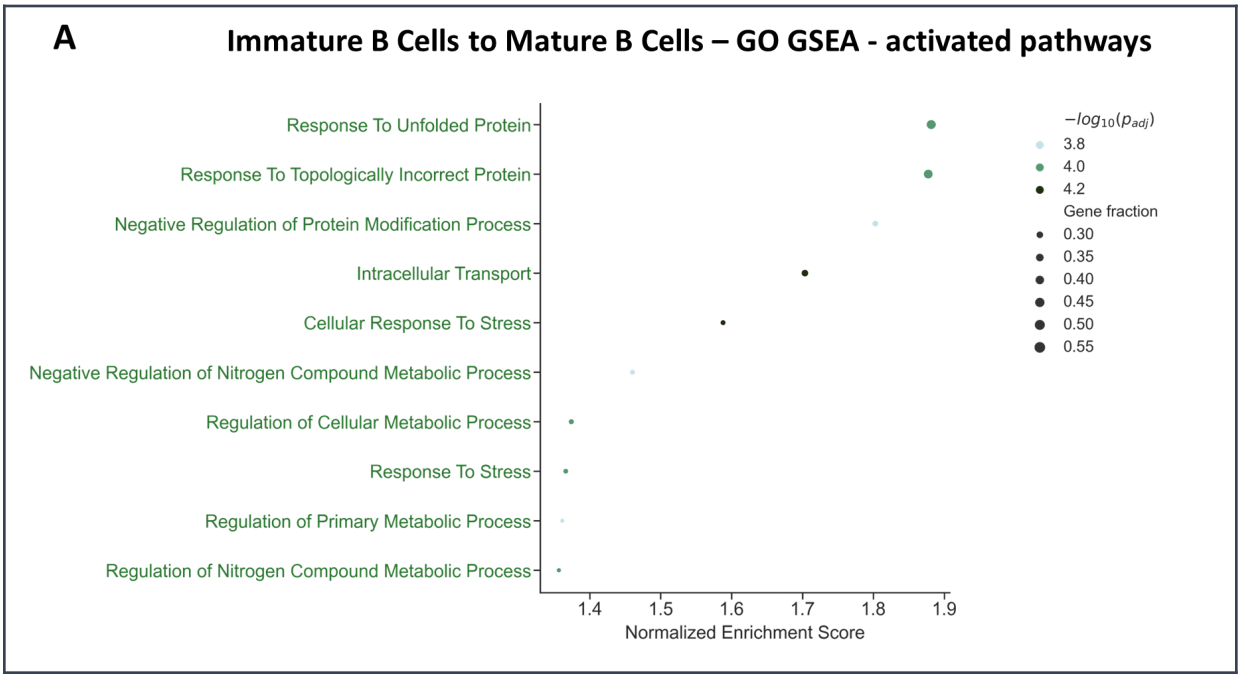


Figure 4. (A) GSEA using the GO gene dataset reveals activation of genes pathways associated with proliferation and regulation of Lymphocyte-cells and suppression of biosynthetic process, based on biolord's counterfactual predictions of the gene expression of Pre-pro B cells shifted from the Yolc sac to the spleen. (B) A Heatmap of mean expression of Pre-pro B cells sampled from the yolk sac at 7 pcw vs the predicted gene expression after shifting the cells each time from the yolk sac to a different target organ. Each row corresponds to the gene expression within a distinct organ, while each column signifies a gene within the enclosed gene set. (C) Taken from Figure 8 from the work of Qilong Li et al. [28]; Illustration of the Docking Interactions involving the Human c-Fos-c-Jun complex and DNA, alongside DHA. DHA exerts dual roles—it not only attaches to the human c-Fos protein, enhancing the interaction by establishing a hydrogen bond with the 158th Arginine residue but also binds to the DNA at the thirty-first adenine position [28].

Continuing on this path, we advanced Pre-pro B cells from the liver at 16 pcw through their developmental stages. This time we focused on a different subset of genes. This set contains 33 genes, categorized as control random genes (10), immune response and B-cell development (15), and cell signaling and regulation genes (8). The latter two groups contain genes recognized as markers for distinct B-cell developmental stages [3,30,31].

Echoing prior findings, the random control group exhibited minimal variation among diverse cell types. In contrast, the other two groups exposed patterns of variation as illustrated in Figure 5B. For example, genes like MME and CD99 exhibited expression limited to early developmental stages, aligning with existing knowledge [3]. This trend mirrors observations in IGKC, specific to mature B cells as part of the immunoglobulin light chain development [32,33].

Furthermore, our analysis unveiled cell type-specific expression, exemplified by the high expression of CCR7 in B1 cells [30]. This distinct pattern underscores the intricate, context-dependent regulatory mechanisms governing gene expression throughout B-cell development.



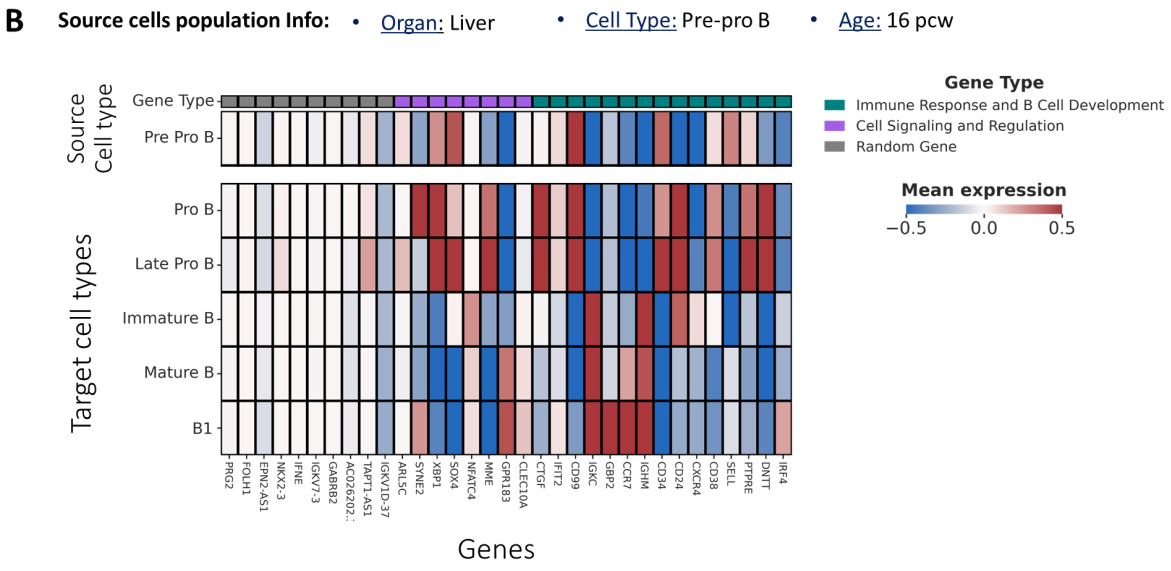


Figure 5. (A) GO-GSEA generated from biolord's predictions to shed light on gene expression transitions from Immature B cells to Mature B cells in the liver at 16 pcw. This analysis unveils genes associated with intracellular transport activation, cellular stress response, and metabolic regulation. (B) A Heatmap of mean expression in Pre-pro B cells sampled from the liver at 16 pcw versus predicted expression after successive shifts through cells' developmental stages. Rows correspond to distinct cell types, while columns represent genes from the enclosed 33-gene set. The set is divided into four color-coded groups: Control random genes (10), Immune Response and B Cell Development (15), and Cell Signaling and Regulation genes (8).

Discussion

The comprehensive exploration of the immune system development and more specifically B cell maturation performed in this study offers valuable insights into the complex interplay of biological processes within the human body. The immune system's development involves a coordinated effort among various organs and cell types, resulting in the establishment of a robust defense mechanism against pathogens. The understanding of these intricacies holds significant potential for advancing medical science, developing innovative immunotherapies, and devising targeted strategies for immune-related disorders.

The use of biolord to disentangle cell attributes offers a novel perspective, revealing distinct clustering patterns based on organ and cell type. Moreover, it allows the manipulation of cell attributes, as demonstrated by the transition of Pre-pro B cells to the spleen, thereby unearthing shifts in gene expression profiles associated with organ identity. Further, the analysis

offers insights into the filtration processes occurring within the spleen's germinal centers. This approach sheds light on organ influence and demonstrates the potential of computational methods to simulate cell migration.

Moreover, analysis of gene trends related to B-cell types uncovered activation pathways associated with cellular stress response and metabolic regulation. These findings highlight the significance of stress response for B cell vitality and provide insights into the intricate regulatory mechanisms governing gene expression during B-cell development.

Throughout this study, the majority of results have aligned with established biological knowledge, reinforcing their validity, and the validity of the novel computational approach. However, an intriguing aspect emerged in certain cases, particularly when analyzing the expression patterns of specific genes in response to shifts in cell type. In these instances, unexpected expression profiles arose that deviated from the current literature. An illustrative example is the gene CD38, which, contrary to our observations, was found to be also expressed in the later stages of B-cell development [33]. This apparent discrepancy doesn't necessarily indicate a limitation in biolord's prediction capabilities; rather, it presented the inherent complexity within the B-cell population, consisting of diverse sub-populations each with unique gene profiles. This insight prompts us to consider a closer look at specific lineage pathways, potentially enabling more accurate predictions by capturing the distinct diversity and variations within cell populations. Further exploration along these lines could unveil deeper insights into the dynamics of gene expression during B-cell development and refine the predictive power of computational models like biolord.

For future research, diverse datasets integration stands as a significant consideration. The inclusion of additional data sources, along with the data used here [1] could further enhance our insights. A promising approach could include integrating more types of single-cell data, such as Ribo-Seq. This would give us a more complete view of which mRNAs are actively being translated. As this technology progresses with improvements in single-cell resolution, exemplified by the research conducted by Ozadam et al [34]. Thus, by incorporating data from various stages of the Central dogma we could better uncover biological processes with the help of biolord.

Another intriguing path to explore with biolord is the comparison of expression signatures across a broader range of cell populations. This approach holds the potential to uncover crucial genetic markers and further our understanding of the underlying mechanisms driving autoimmune conditions.

In conclusion, this study harnesses computational tools like biolord to unravel hidden biological processes, enhancing our understanding of the immune system dynamics and showcasing the potential of computational methods in simulating cell migration and cell state shifts across development.

Methods

Data gathering and pre-processing

The data for this study originates from Suo et al. [1] and was accessed through cellxgene: Lymphoid cells. The dataset encompasses a comprehensive single-cell atlas, spanning the developmental spectrum of human immune cells across prenatal hematopoietic, lymphoid, and nonlymphoid peripheral organs. This extensive dataset comprises over 900,000 cells, further categorized into more than 100 distinct cell states. In line with our focus on B-cells, we initially excluded other cell types leaving us with 64K cells. Subsequently, we carried out data preprocessing using standard Python libraries, including Scanpy and Anndata. Key preprocessing steps involved normalizing each cell based on the total counts of all genes, log-scaling the data, and filtering out genes with non-highly variable expressions. All the code for this project is available as open source in biolord_immune_bcells.

biolord (biological representation disentanglement)

Biolord is a deep generative framework for disentangling known and unknown attributes in single-cell data recently presented by Piran et al [9]. This tool uses the assumption that the single-cell data (e.g. gene expression) can be disentangled into latent representations, consisting of multiple layers (spaces) each associated with a single known attribute, marked as Z_y, and an additional space for all the unknown attributes in the data, marked as Z_u.

As seen in Figure 6, biolord takes as input the dataset which comprises of the gene expression and the labeled attributes and constructs a decomposed latent space using a set of dedicated neural networks [9]. The output of each sub-network is either Z_y the latent space representing each attribute or Z_u the latent space continuing all the unknown attributes. We note that there is no attempt to distinguish or identify these attributes. The latent space embedding maps each sample to a latent code. Next, a generator is defined, marked as $G_{\theta,}$ which takes as input the decomposed latent space and outputs predictions for the gene expression.

The disentangled representation task is completed by inducing information constraints on the joint loss function; the loss attempts to maximize the accuracy of the reconstruction (enforcing completeness) while minimizing the information encoded in the unknown attributes (limiting its capacity) [9]. The decomposed latent space allows the study of the different attributes and their inner structure independently. Moreover, using the generative part of the model we can perform data manipulation, such as querying how a cell's expression would have changed if it was sampled at a different time point or a different tissue [9]. More generally, the framework allows studying gene trends as a function of cell state (Figure 3 A and B). Biolord was implemented using the scvi-tools library [35] and is available as open-source software at https://github.com/nitzanlab/biolord.

In this project, we trained biolord over the B-cells dataset [1] with the purpose of receiving a latent disentangled representation of the single-cell data including the known attributes: the cell type, the source organ, and the age and sex of the donor, referred to as categorical attributes. Hyperparameters such as the autoencoder width and depth were selected after a finetuning procedure using wandb [36]. All the code for this project is available in biolord_immune_bcells.

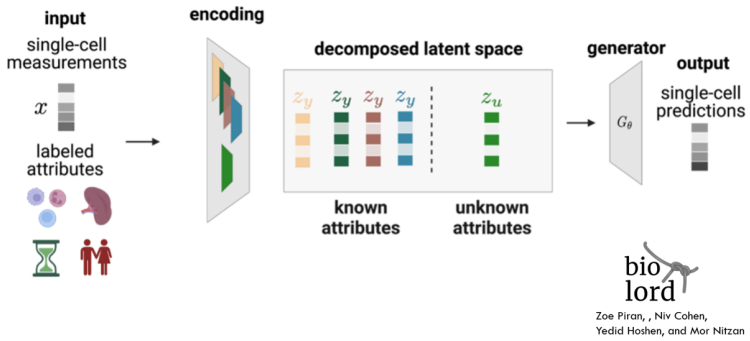


Figure 6. biolord - biological representation disentanglement. A deep generative framework for disentangling known and unknown attributes in single-cell data. This tool takes as input the single-cell measurements and labeled attributes which may be categorical (discrete; e.g. cell type) or ordered (continuous; e.g. age). Then using an architecture of encoder-decoder we obtain a disentangled representation for all (known and unknown) attributes in the data. On top of these, biolord learns a generator, which maps the representations of the known and unknown attributes into observable single-cell data [9].

References

1. Suo C, Dann E, Goh I, Jardine L, Kleshchevnikov V, Park J-E, et al. Mapping the developing human immune system across organs. Science. 2022 [cited 6 Jun 2023]. doi:10.1126/science.abo0510

2. Melchers F. Checkpoints that control B cell development. J Clin Invest. 2015;125: 2203–2210. doi:10.1172/JCI78083

3. Hystad ME, Myklebust JH, Bø TH, Sivertsen EA, Rian E, Forfang L, et al. Characterization of early stages of human B cell development by gene expression profiling. J Immunol. 2007;179: 3662–3671. doi:10.4049/jimmunol.179.6.3662

4. Wang Y, Liu J, Burrows PD, Wang J-Y. B Cell Development and Maturation. In: Wang J-Y, editor. B Cells in Immunity and Tolerance. Singapore: Springer Singapore; 2020. pp. 1–22. doi:10.1007/978-981-15-3532-1_1

5. Sandel PC, Monroe JG. Negative selection of immature B cells by receptor editing or deletion is determined by site of antigen encounter. Immunity. 1999;10: 289–299. doi:10.1016/s1074-7613(00)80029-1

6. King LB, Monroe JG. Immunobiology of the immature B cell: plasticity in the B-cell antigen receptor-induced response fine tunes negative selection. Immunol Rev. 2000;176: 86–104. doi:10.1034/j.1600-065x.2000.00609.x

7. Chen H, Ye F, Guo G. Revolutionizing immunology with single-cell RNA sequencing. Cell Mol Immunol. 2019;16: 242–249. doi:10.1038/s41423-019-0214-4

8. Kopf A, Claassen M. Latent representation learning in biology and translational medicine. Patterns (N Y). 2021;2: 100198. doi:10.1016/j.patter.2021.100198

9. Wieczorek A, Wieser M, Murezzan D, Roth V. Learning Sparse Latent Representations with the Deep Copula Information Bottleneck. arXiv [stat.ML]. 2018. Available: http://arxiv.org/abs/1804.06216

10. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut Learning in Deep Neural Networks. 2020 [cited 22 Aug 2023]. doi:10.1038/s42256-020-00257-z

11. Wang X, Chen H, Tang S 'ao, Wu Z, Zhu W. Disentangled Representation Learning. 2022. Available: http://arxiv.org/abs/2211.11695

12. Piran Z, Cohen N, Hoshen Y, Nitzan M. Biological representation disentanglement of single-cell data. bioRxiv. 2023. p. 2023.03.05.531195. doi:10.1101/2023.03.05.531195

13. Locatello F, Bauer S, Lucic M, Rätsch G, Gelly S, Schölkopf B, et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. 2018. Available: http://arxiv.org/abs/1811.12359

14. Locatello F, Tschannen M, Bauer S, Rätsch G, Schölkopf B, Bachem O. Disentangling Factors of Variation Using Few Labels. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1905.01258

15. Chen Y, Mancini M, Zhu X, Akata Z. Semi-Supervised and Unsupervised Deep Visual Learning: A Survey. IEEE Trans Pattern Anal Mach Intell. 2022;PP. doi:10.1109/TPAMI.2022.3201576

16. Burgess CP, Matthey L, Watters N, Kabra R, Higgins I, Botvinick M, et al. MONet: Unsupervised Scene Decomposition and Representation. 2019. Available: http://arxiv.org/abs/1901.11390

17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

18. Loder F, Mutschler B, Ray RJ, Paige CJ, Sideras P, Torres R, et al. B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. J Exp Med. 1999;190: 75–89. doi:10.1084/jem.190.1.75

19. Qiao Y, Hu H, Zhao Y, Jin M, Yang D, Yin J, et al. Benzene induces spleen injury through the B cell receptor signaling pathway. Ecotoxicol Environ Saf. 2023;257: 114924. doi:10.1016/j.ecoenv.2023.114924

20. Zhao L, Liu L, Guo B, Zhu B. Regulation of adaptive immune responses by guiding cell movements in the spleen. Front Microbiol. 2015;6: 645. doi:10.3389/fmicb.2015.00645

21. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018. doi:10.1038/nbt.4314

22. Kach J, Sethakorn N, Dulin NO. A finer tuning of G-protein signaling through regulated control of RGS proteins. Am J Physiol Heart Circ Physiol. 2012;303: H19–35. doi:10.1152/ajpheart.00764.2011

23. Ansel KM, Ngo VN, Hyman PL, Luther SA, Förster R, Sedgwick JD, et al. A chemokine-driven positive

feedback loop organizes lymphoid follicles. Nature. 2000;406: 309–314. doi:10.1038/35018581

24. Denoeud J, Moser M. Role of CD27/CD70 pathway of activation in immunity and tolerance. J Leukoc Biol. 2011;89: 195–203. doi:10.1189/jlb.0610351

25. Regulation of immune-related diseases by multiple factors of chromatin, exosomes, microparticles, vaccines, oxidative stress, dormancy, protein quality control, inflammation and microenvironment: a meeting report of 2017 International Workshop of the Chinese Academy of Medical Sciences (CAMS) Initiative for Innovative Medicine on Tumor Immunology. Yao Xue Xue Bao. 2017;7: 532–540. doi:10.1016/j.apsb.2017.06.007

26. Zemskov AM, Zemskov VM. An integral concept of regulating immune homeostasis. Biology Bulletin Reviews. 2015;4: 467–476. doi:10.1134/S2079086414060085

27. Karakaslar EO, Katiyar N, Hasham M, Youn A, Sharma S, Chung C-H, et al. Transcriptional activation of Jun and Fos members of the AP-1 complex is a conserved signature of immune aging that contributes to inflammaging. Aging Cell. 2023;22: e13792. doi:10.1111/acel.13792

28. Li Q, Jiang N, Zhang Y, Liu Y, Su Z, Yuan Q, et al. Dihydroartemisinin imposes positive and negative regulation on Treg and plasma cells via direct interaction and activation of c-Fos. Commun Biol. 2023;6: 52. doi:10.1038/s42003-023-04454-5

29. Inada K, Okada S, Phuchareon J, Hatano M, Sugimoto T, Moriya H, et al. c-Fos induces apoptosis in germinal center B cells. J Immunol. 1998;161: 3853–3861. Available: https://www.ncbi.nlm.nih.gov/pubmed/9780150

30. Luo Y, Wang J, Li K, Li M, Xu S, Liu X, et al. Single-cell genomics identifies distinct B1 cell developmental pathways and reveals aging-related changes in the B-cell receptor repertoire. Cell Biosci. 2022;12: 57. doi:10.1186/s13578-022-00795-6

31. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. J Allergy Clin Immunol. 2013;131: 959–971. doi:10.1016/j.jaci.2013.01.046

32. Lee RD, Munro SA, Knutson TP, LaRue RS, Heltemes-Harris LM, Farrar MA. Single-cell analysis of developing B cells reveals dynamic gene expression networks that govern B cell development and transformation. bioRxiv. 2020. p. 2020.06.30.178301. doi:10.1101/2020.06.30.178301

33. Morgan D, Tergaonkar V. Unraveling B cell trajectories at single cell resolution. Trends Immunol. 2022;43: 210–229. doi:10.1016/j.it.2022.01.003

34. Ozadam H, Tonn T, Han CM, Segura A, Hoskins I, Rao S, et al. Single-cell quantification of ribosome occupancy in early mouse development. Nature. 2023;618: 1057–1064. doi:10.1038/s41586-023-06228-9

35. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022;40: 163–166. doi:10.1038/s41587-021-01206-w

36. Biewald L. Experiment tracking with weights and biases. Software available from wandb. com. 2020. URL https://www wandb com. 2020.