PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Pseudocell Tracer—A method for inferring dynamic trajectories using scRNAseq and its application to B cells undergoing immunoglobulin class switch recombination

Derek Reiman[1☯], Godhev Kumar Manakkat Vijay[2☯], Heping Xu[3,4☯], Andrew Sonin[5], Dianyu Chen[3,4], Nathan Salomonis[6], Harinder Singh[2]*, Aly A. Khan[7]*

1 University of Illinois at Chicago, Department of Bioengineering, Chicago, Illinois, United States of America, **2** University of Pittsburgh, Center for Systems Immunology, Departments of Immunology and Computational and Systems Biology, Pittsburgh, Pennsylvania, United States of America, **3** Key Laboratory of Growth Regulation and Translation Research of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China, **4** Institute of Biology, Westlake Institute for Advanced Study, Hangzhou, Zhejiang Province, China, **5** Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia, **6** Division of Biomedical Informatics, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States of America, **7** University of Chicago, Department of Pathology, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
* harinder@pitt.edu (HS); aakhan@uchicago.edu (AAK)

Citation: Reiman D, Manakkat Vijay GK, Xu H, Sonin A, Chen D, Salomonis N, et al. (2021) Pseudocell Tracer—A method for inferring dynamic trajectories using scRNAseq and its application to B cells undergoing immunoglobulin class switch recombination. PLoS Comput Biol 17(5): e1008094. https://doi.org/10.1371/journal.pcbi.1008094

Editor: Jason M. Haugh, North Carolina State University, UNITED STATES

Received: June 24, 2020

Accepted: March 30, 2021

Published: May 3, 2021

Data Availability Statement: All raw single cell RNA-seq data from this work is submitted to the GEO repository: GSE171867. Software code used in generating the results is described above in detail and on GitHub: https://github.com/akds/pseudocell.

Funding: The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research (AK), UPMC-ITTC

Abstract

Single cell RNA sequencing (scRNAseq) can be used to infer a temporal ordering of cellular states. Current methods for the inference of cellular trajectories rely on unbiased dimensionality reduction techniques. However, such biologically agnostic ordering can prove difficult for modeling complex developmental or differentiation processes. The cellular heterogeneity of dynamic biological compartments can result in sparse sampling of key intermediate cell states. To overcome these limitations, we develop a supervised machine learning framework, called Pseudocell Tracer, which infers trajectories in pseudospace rather than in pseudotime. The method uses a supervised encoder, trained with adjacent biological information, to project scRNAseq data into a low-dimensional manifold that maps the transcriptional states a cell can occupy. Then a generative adversarial network (GAN) is used to simulate pesudocells at regular intervals along a virtual cell-state axis. We demonstrate the utility of Pseudocell Tracer by modeling B cells undergoing immunoglobulin class switch recombination (CSR) during a prototypic antigen-induced antibody response. Our results revealed an ordering of key transcription factors regulating CSR to the IgG1 isotype, including the concomitant expression of *Nfkb1* and *Stat6* prior to the upregulation of *Bach2* expression. Furthermore, the expression dynamics of genes encoding cytokine receptors suggest a poised IL-4 signaling state that preceeds CSR to the IgG1 isotype.

initiative (HS), and National Natural Science Foundation of China (grant31970842; HX). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

In the past decade advances in computing and single cell sequencing technologies have ushered in a new era of discovery in biology and medicine. However, the analysis of single cell data remains challenging, especially when analyzing heterogeneous cellular compartments with complex dynamics. This scenario is especially pronounced in dynamic immune responses of innate and adaptive immune cells. Existing computational tools typically analyze scRNAseq datasets without reference to any of the underlying biology of the system that generates the data. We reason that use of prior knowledge of the system can aid in the extraction of obscured information from scRNAseq datasets. We introduce a framework, Pseudocell Tracer, which takes advantage of validated biological knowledge to guide the inference of cellular trajectories. We apply and validate Pseudocell Tracer by scRNAseq analysis of antigen-specific B cells undergoing immunoglobulin class switch recombination during an antigen-induced humoral immune response. This framework is potentially applicable to single cell data from many other fields with complex dynamics.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Single-cell RNA sequencing (scRNAseq) has emerged as a dominant tool for analyzing the transcriptional states of individual cells in diverse biological contexts [1,2]. Computational analyses of scRNAseq datasets have enabled rigorous delineation of known cellular identities as well as the discovery of novel cell types [3]. Such datasets have also been used to infer a temporal ordering of dynamic cellular states or cellular trajectories [4]. For example, the field of immunology has benefited significantly from the adoption of scRNAseq in order to characterize cellular states in the context of development and differentiation of distinct innate and adaptive lineages [5–7], including responses to various perturbations [8–10], as well as in immune system diseases [11–13]. Despite tremendous progress, the inference of cellular trajectories from scRNAseq datasets remains challenging when analyzing heterogeneous cellular compartments with complex dynamics.

Current computational methods for cellular trajectory inference rely on two crucial steps. In the first step, dimensionality reduction techniques [14], such as PCA [15], ICA [16], and UMAP [17], are used to project and visualize single cells based on their gene expression profiles in low dimensional space (Fig 1A, left). While single cell transcriptional profiles have high dimensionality due to the thousands of genes profiled, their intrinsic dimensionalities are typically much lower. Gene expression during a biological process can be directed by small combinations of transcription factors that regulate large gene modules in a time-dependent manner. Thus, unsupervised low dimensional projections can reveal salient temporal structure in large-scale scRNAseq datasets, especially when a dominant transcriptional regulatory program directs the biological process. In the second step of trajectory inference, pathfinding algorithms, such as minimum spanning trees [15,16] or k-nearest neighbor graphs [17,18], are utilized for inferring an ordering of the empirically observed cells in the low dimensional space (Fig 1A, center). The cells ordered in the inferred trajectory are typically mapped to a virtual temporal axis called "pseudotime", which is bounded by two cells representing the start and

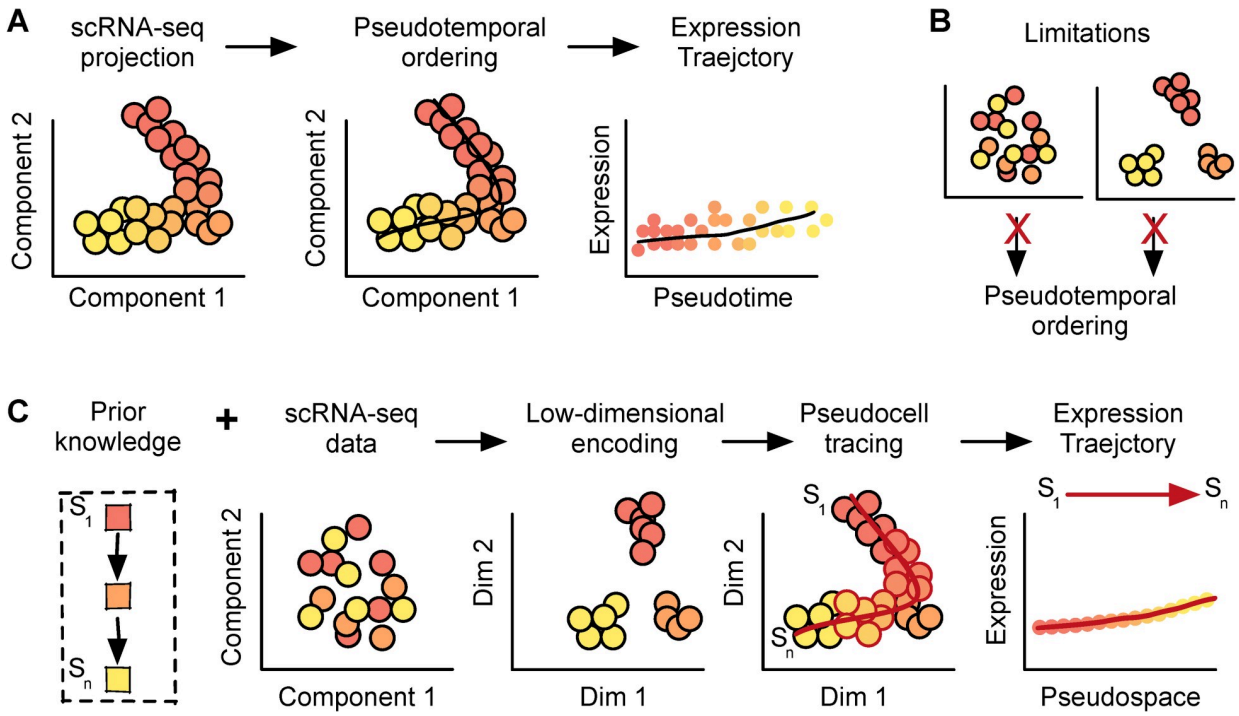

Fig 1. Pseudocell Tracer, a framework for modeling cellular trajectories in complex systems. (A) An overview of pseudotime trajectory inference. (B) Some scenarios that may obstruct pseudotime ordering. (C) Pseudocell Tracer. Given some prior knowledge about a model system, we aim to predict expression trajectories by generating pesudocells at regular intervals along a virtual cell-state axis, even though such cells may be sparsely captured in single cell profiling data.

https://doi.org/10.1371/journal.pcbi.1008094.g001

end of the cellular trajectory. Gene expression abundances from the original high dimensional profiles can then be plotted along a pseudotime coordinate to display their changes along the inferred trajectory (Fig 1A, right).

Heterogeneous cellular compartments with complex temporal dynamics can present unique hurdles to trajectory inference. We consider two scenarios that are common within such complex systems and that limit the use of standard inference methods. In the first scenario, cells utilizing concurrent transcriptional regulatory modules, such as those controlling cell cycle, metabolism and differentiation, may not reveal the developmental trajectory of interest along any particular axis using unsupervised dimensionality reduction techniques [19] (Fig 1B, left). Dimensionality reduction works by minimizing (or maximizing) some global statistical measure on the gene expression profiles, such as percent variance explained in each orthogonal dimension of PCA. Thus, there is no guarantee that any single unsupervised dimensionality reduction technique can uncover a specific temporal pattern of interest. In the second scenario, all transitioning cellular states along a given trajectory may not be equally populated, resulting in greater capture of some cell states and sparse capture of other states (Fig 1B, right). This non-uniform sampling occurs when cells do not follow a constant rate of development or differentiation during a time-dependent process. Consequently, this can result in the lack of an observable continuum of cell states or temporal structure in low dimensional space, hindering cell ordering. Thus, these scenarios illustrate some of the key hurdles to trajectory inference for complex cellular compartments.

Existing computational tools for analyzing scRNAseq datasets typically do so without reference to any of the underlying biological guideposts of the system used to generate the data. We hypothesized that by developing algorithms that take advantage of validated prior biological

knowledge, we could extract otherwise unresolvable trajectories from scRNAseq datasets, especially when analyzing heterogeneous cellular compartments with complex dynamics. To test this hypothesis, we develop a supervised machine learning framework, called Pseudocell Tracer, that enables modeling of cellular trajectories in complex dynamical systems. We accomplish supervision by harnessing adjacent information about the underlying biological process. In most biological systems, there is some prior validated knowledge about the underlying cellular states, directionality and dynamics of the process, which can be integrated into a computational model. For example, cellular differentiation is often tracked by the level of expression of one or more specific markers such as cell surface proteins or regulators i.e., transcription factors. The expression level of such markers or regulators can reflect a "developmental clock" and therefore serve as an estimate of the progression between the progenitor and terminal differentiation state in a complex biological process. We refer to this kind of prior knowledge as adjacent biological information.

To implement Pseudocell Tracer we harnessed recent advances in deep generative modeling. Generative adversarial networks (GANs) can learn a latent space from which to simulate gene expression profiles of cells that are indistinguishable from a distribution of real cells [20]. In particular, it has been previously proposed that the interpolation of cells in the latent space may be a means for simulating pseudocells along some cellular trajectory [21]. Notably, GANs cannot directly shape the latent space, for example, to reflect prior knowledge about a complex cellular compartment. However, autoencoders can learn latent spaces for scRNAseq that satisfy specific biological constraints [22]. The integration of such models along with the use of adjacent biological information to supervise their training has not received significant attention. Pseudocell Tracer integrates such models and generates pseudocells along specific cellular trajectories in a stepwise process. First, Pseudocell Tracer uses an encoder (Fig 1C, left) to map out a low-dimensional manifold that describes the transcriptional states a cell can occupy, while remaining faithful to adjacent biological information. Then, the framework uses a generator (Fig 1C, center) to simulate pesudocells at regular intervals in the latent space by using the same adjacent biological information as a guide. Finally, these pseudocells are subjected to a decoder (Fig 1C, right) to observe gene expression dynamics along the trajectory and provide novel insights into the underlying regulatory mechanisms. In contrast to pseudotime inference methods which seek to order empirically observed cell states, Pseudocell Tracer instead generates (pseudo) cells along a defined cell state-space interval. Thus, Pseudocell Tracer infers trajectories in "pseudospace" rather than in "pseudotime".

We apply Pseudocell Tracer to the process of somatic DNA recombination that B cells of the immune system undergo upon antigen encounter. The process termed class switch recombination (CSR) results in exchange of the constant region (M isotype) of the immunoglobulin heavy chain protein (IgH) to one of several other isotypes thereby generating IgG, IgA and IgE antibody expressing cells [23,24]. B cells that switch their IgM locus to one of the other Ig isotypes, via DNA recombination, can be viewed as moving down distinct cellular trajectories since different cytokine signals and transcription factors have been shown to promote specific types of isotype switching. An understanding of the timing and expression of the various signaling components and transcription factors associated with distinct CSR trajectories remains to be thoroughly explored and has not been analyzed in vivo by single-cell transcriptional profiling. Using an scRNAseq dataset generated in the context of a prototypic antigen-specific B cell response we demonstrate that standard trajectory inference methods fail to assemble appropriate CSR trajectories. Instead, Pseudocell Tracer trained with adjacent information in the form of relative expression of isotype-specific transcripts enhances both dimensionality reduction and trajectory inference. In so doing it reveals the relative timing and orchestration of key cytokine receptors and transcription factors regulating a particular CSR trajectory.

Results

Experimental system and the scRNAseq dataset

Mouse B cell responses to the model antigen 4-hydroxy-3-nitrophenylacetyl-keyhole limpet hemocyanin (NP-KLH) have been used to reveal fundamental principles underlying antibody isotype switching and affinity maturation [25,26]. To analyze the dynamic transcriptional states of activated B cells we performed scRNAseq on NP-specific germinal center B cells at the peak of the response (day 14) (Fig 2A). We reasoned that in B cells undergoing CSR, Ig transcripts encoding the M isotype would diminish whereas those encoding the switched isotype would increase.

We calculated the relative isotype expression within a B cell by dividing the log2(TPM + 1) expression of each distinct isotype transcript (Ighm, Ighg1, Ighg2b, Ighg2c, Ighg3, Igha, Ighd, and Ighe) with the cumulative expression of all isotypes. Hierarchical clustering of relative isotype expression revealed 7 clusters of B cells, of which 5 clusters were dominated by a single isotype and reflected cells that had undergone CSR (Fig 2B). As expected for the immunization

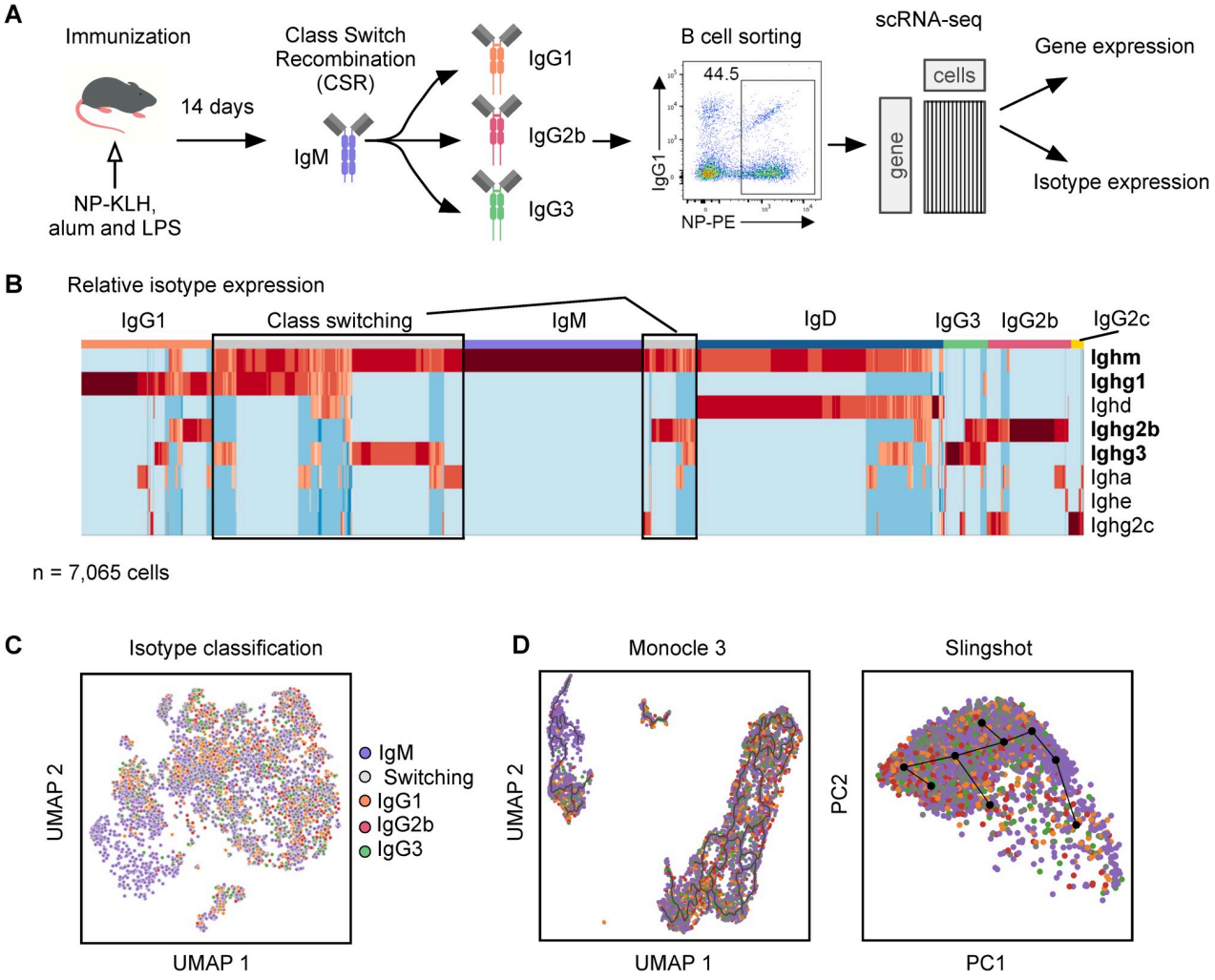

Fig 2. Class switch recombination process. (A) Overview of experimental system. (B) Relative isotype expression for all cells. N = 7,065. All isotype expressions sum to one for a given cell. (C) UMAP of RNA-seq data, colored by isotype. (D) Output from Moncole3 and Slingshot.

https://doi.org/10.1371/journal.pcbi.1008094.g002

conditions noted above, IgM expressing B cells primarily switched their isotypes to IgG1, IgG2b, or IgG3 (Stavnezer et al., 2008). Notably, the clustering based on relative isotype expression values revealed a group of transitioning cells that were undergoing CSR from IgM to other isotypes, including IgG1, IgG2b, and IgG3. Since the relative isotype expression is expected to monotonically track the progression of CSR, we reasoned that it would represent suitable information embedded within the dataset to enable deconvolution of the distinct CSR cellular trajectories within the transitioning B cells.

We visualized the scRNAseq data using uniform manifold approximation and projection (UMAP) and observed that it failed to distinguish cells on the basis of their isotype identities (Fig 2C). We therefore evaluated the utility of existing computational pipelines, in particular, Moncole3 [27] and Slingshot [15] for delineating CSR trajectories within our dataset. These methods utilize a variety of dimensionality reduction techniques and temporal ordering methods. However, none of the methods recovered a coherent CSR trajectory that delineated a path from, for example, IgM to IgG1 (Fig 2D). In the case of Monocle3, the low dimensional projection of single cell data by UMAP was neither able to distinguish stably switched cells by their isotype nor cluster cells presumptively undergoing CSR. Similarly, the Slingshot representation failed to distinguish cells on the basis of their CSR trajectories.

We hypothesized that the failure of these unsupervised dimensionality reduction techniques to uncover the CSR trajectories was due to other dominant dynamic gene expression programs in germinal center B cells, particularly involving the cell cycle. To evaluate this hypothesis, we analyzed the expression patterns of cell cycle regulators, which revealed clustering of cells based on their cell cycle phase (S1 Fig). Taken together, while our directed analysis of the scRNAseq dataset revealed a minor cluster of transitioning B cells that were undergoing CSR to different isotypes, existing unsupervised methods were unable to reveal such cells as a distinct cluster(s) and therefore temporally order their pertinent trajectories.

Overview of pseudocell tracer

The core of the Pseudocell Tracer framework (Fig 3A) is based on the following two components that are used successively: (1) a supervised autoencoder to perform dimensionality reduction and (2) a conditional generative adversarial network (CGAN) to generate hypothetical cell states or pseudocells. The main difference between an unsupervised and supervised autoencoder is the additional information provided to facilitate learning a low dimensional projection (Fig 3A, left). Both unsupervised and supervised autoencoders function to encode high-dimensional data into a low-dimensional latent space. However, the supervised autoencoder aims to specifically learn an encoder that transforms the scRNAseq data into a latent space that conforms to the adjacent information, relating to a specific biological context or process. In the context of modeling CSR, the latent space is shaped by relative expression of the different Ig constant region transcripts. Thus, individual B cells with similar relative isotype expression profiles will have similar latent encodings. The architecture of the supervised autoencoder contains both an encoder and decoder (S2 Fig). The encoder functions to project high-dimensional data into a low-dimensional latent space, which is shaped by the adjacent biological information (Fig 3B, top), and the decoder functions to reverse the low dimensional encoding of cells from the latent space (Fig 3B, bottom). When combined together, the encoder performs dimensionality reduction, while the decoder generates from the latent space a reconstruction as close as possible to the observed input data.

Visualization of the latent space for the scRNAseq data revealed specific clustering of cells by their dominant isotype (Fig 3B, top). To characterize the robustness of the model on new or held out data, we evaluated the supervised autoencoder using 10-fold cross-validation (Fig

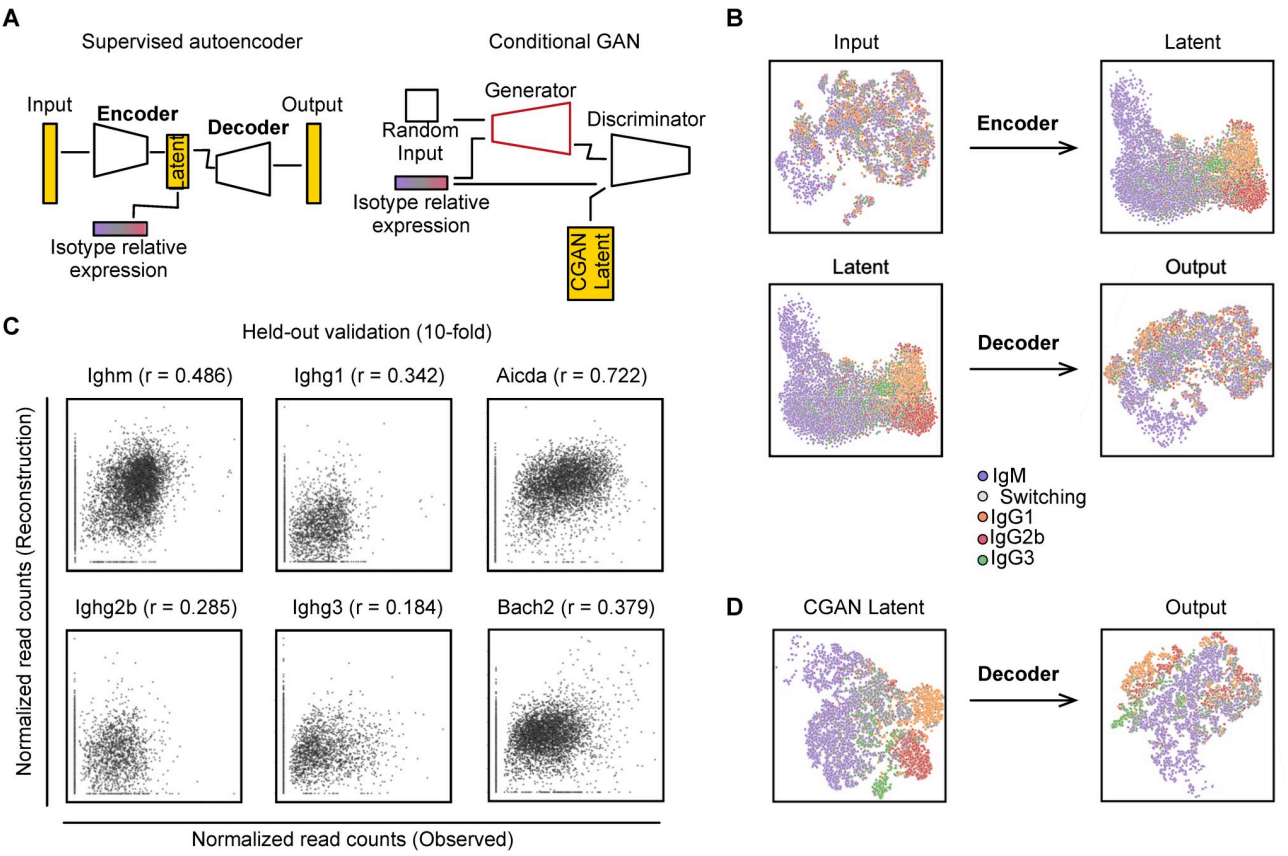

Fig 3. Pseudocell Tracer efficiently integrates adjacent biological information and accurately simulates gene expression profiles in pseudocells. (A) Overview of neural network model combining a supervised autoencoder with a conditional GAN. (B) UMAP visualization of the input and output used in the supervised autoencoder; encoder (top) and decoder (bottom). (C) Scatter plot between observed and predicted expression values on held out cells. r denotes Pearson correlation between ground truth and predicted values. Isotype expression (left) and example CSR genes (right). (D) UMAP visualization applied to cGAN prediction and subsequent output from decoder.

https://doi.org/10.1371/journal.pcbi.1008094.g003

3C). For each partition, 90% of the data was used for training and 10% was set aside as a blind test. For training, an additional 10% of the training set was used for early stopping (see Methods). Once training finished, the test set was then encoded and decoded. Scatterplots of the held-out test predictions were generated for IgM and IgG isotypes, demonstrating high correlation between predicted and observed gene expression (Fig 3C). Notably, other factors associated with CSR, such as *Aicda* and *Bach2*, demonstrated high correlation. Next, we sought to evaluate how sensitive the model is to varying sample sizes. We examined correlation values between predicted and observed expression of IgM and IgG isotypes using 500, 1000, and 5000 cells (S1 Table). Training with larger numbers of cells produced a better latent representation of the single cell data and improved modelling of gene expression profiles. Taken together, the supervised autoencoder successfully learned both an encoder for dimensionality reduction informed by relative isotype expression and a decoder for mapping low dimensional encodings back to full transcriptional expression profiles.

In the second step of our approach, we trained a CGAN to simulate pseudocells (S3 Fig). Notably, the inference procedure for the generative model is performed in the latent space that is learned by the autoencoder. The use of the low-dimensional latent space is necessitated by the instability of GANs in high dimensional settings. Importantly, any CGAN simulation in low dimensional space can be easily mapped to the input space using the decoder and generate

high dimensional full transcriptional profiles for individual cells. The main difference between a GAN and a CGAN is the conditional information associated with the generator and discriminator (Fig 3A, right). Both consist of two neural networks competing against each other such that one network, called the generator, seeks to produce realistic output data from a random input vector, and the other network, called the discriminator, is tasked with discriminating between the real and generated data. Importantly, the CGAN conditions the inference of both the generator and the discriminator on adjacent information. In the context of modeling CSR, the generator aims to simulate realistic latent encodings of cells that are conditioned on relative isotype expression profiles, in the same manner as the autoencoder in the first step. Thus, after fully training the CGAN, latent encodings for pesudocells can be simulated using the generator based on their relative isotype expression profiles. These latent encodings can then be subjected to the decoder utilized in the previous step to generate high dimensional transcriptional profiles of hypothetical cells that conform to the input data as well as the adjacent information that represents a key biological prior(s).

To qualitatively evaluate the representation learned by the generator, relative isotype expression and corresponding low-dimensional encodings from the scRNAseq data were used to train the CGAN. The CGAN model was trained to equilibrium. Notably, only the discriminator directly observes low-dimensional encodings of the expression data while the generator improves its simulations through interaction with the discriminator. Pseudocells were generated for the observed relative isotype expressions, subsequently decoded, and visualized (Fig 3D). Visualization of the CGAN latent space reproduced specific clustering of cells by their dominant isotype (Fig 3B) and subsequent decoding reconstructed a complex and heterogeneous B cell landscape qualitatively similar to the real scRNAseq data. Thus through sequential application of an autoencoder and a CGAN both conditioned with prior biological information, *Pseudocell Tracer* provided a supervised framework for generation of hypothetical B cells undergoing CSR.

Pseudocell tracing the CSR process

We define a B cell IgH isotype trajectory based on a cellular progression from the IgM to an alternate IgH isotype. To demonstrate the utility of *Pseudocell Tracer* in inferring cellular trajectories that can be overwhelmed in complex and heterogeneous cellular compartments, we modeled the IgM to IgG1 class switch recombination process. First, we simulated a relative isotype expression profile with IgM at 100% and all other isotypes at 0%. For each cell-state increment along the IgG1 trajectory, we reduced the relative abundance of IgM by 1% and increased the relative abundance of IgG1 by 1%. We continued generating relative isotype expression profiles until IgG1 reached 100% and IgM reached 0% (Fig 4A). Overall, we simulated 101 points along the IgG1 trajectory. We then generated 100 latent encodings for each point using the previously trained CGAN in order to estimate a 95% confidence interval. Finally, we used the previously trained decoder to convert each latent encoding to a full transcriptional expression profile, resulting in 10,100 pseudocells which traced the progression from IgM to the IgG1 state within the trajectory.

To determine if the pseudocell tracing of the IgG1 trajectory was consistent with known experimental findings, we examined the transcriptional dynamics of *Aicda* gene expression in relation to Ighm and Ighg1 transcripts. *Aicda* encodes the activation induced cytidine deaminase (AID) which is a direct mediator of the intrachromosomal IgH recombination events in B cells that result in CSR. We plotted the various relative expression profiles (z-score transformed) across the IgM to IgG1 pseudocell tracing (Fig 4A). As expected, we observed decreased expression of Ighm and increased expression of Ighg1 transcripts as pseudocells

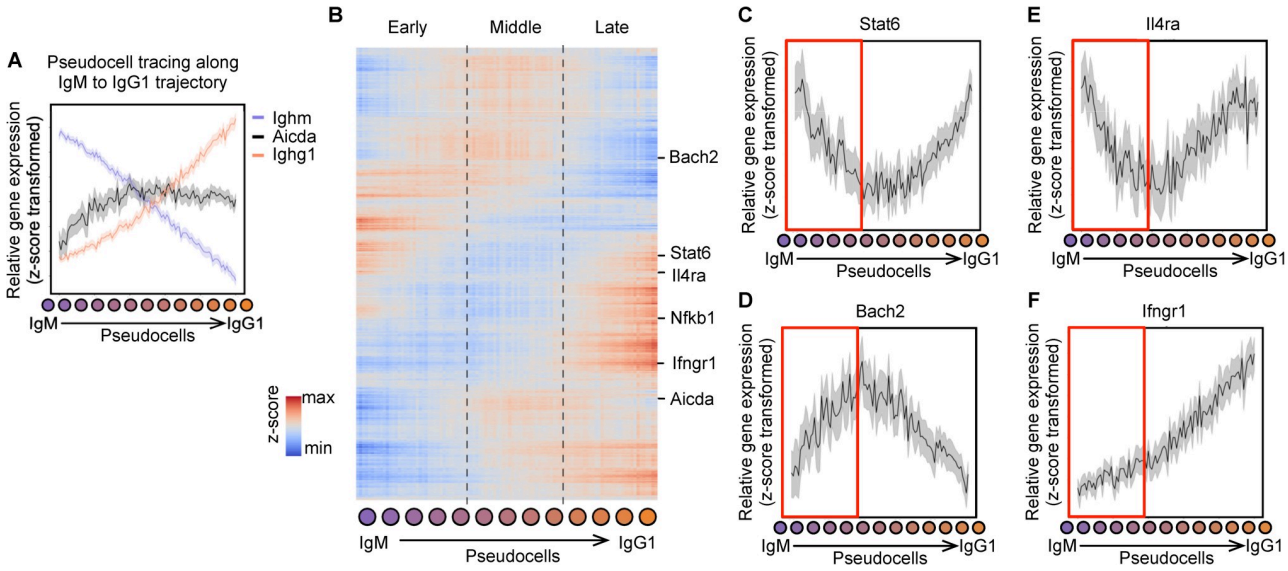

Fig 4. Pseudocell Tracer models IgG1 class switching process. (A) Pseudocells generated along the IgM to IgG1 axis. Plot of relative expression of *Aicda*, *Ighm* and *Ighg1* along the IgM to IgG1 axis, where solid line indicates average expression and shading indicates 95% confidence interval. (B) Associative clustering of genes during CSR. Regions of early (left), middle (center), and late (right) transcriptional dynamics are depicted. Plots of relative expression for key genes with specific dynamics, including (C) Stat6, (D) Bach2, (E) Il4ra, and (F) Ifngr1.

https://doi.org/10.1371/journal.pcbi.1008094.g004

progressed from an IgM cell state to an IgG1 cell state. The *Aicda* transcript profile revealed its increased expression within the CSR trajectory preceeding the inflection point of IgM and IgG1 transcripts. Despite a smaller number of IgG2 and IgG3 cells captured, similar relative expression profiles for *Ighg2*, *Ighg3* and *Aicda* were observed for other IgG trajectories (S4 Fig). The inferred expression profile of *Aicda* suggests a model where its initial levels in antigen-induced GC B cells are likely sufficient for promoting CSR. The increased levels at later timepoints in the CSR trajectory may function in promoting somatic hypermutation (SHM), another key molecular process required for affinity maturation in germinal centers that is directly mediated by AID (see Discussion).

To explore the regulatory underpinnings of IgG1 isotype switching, we assembled a comprehensive view of the gene expression dynamics across the IgG1 trajectory. We focused our analysis on dynamically expressed genes by selecting the top 6,500 most variable genes. Heatmap visualization of the relative expression profiles across the IgM to IgG1 pseudocell tracing revealed 3 granular transcriptional phases associated with this CSR trajectory, designated early, middle, and late (Fig 4B).

We characterized the dynamics of several key transcription factors that are implicated in regulating CSR within the phases. Genes expressed within the early transcriptional phase included *Nfkb1* and *Stat6* (Fig 4B and 4C). *Nfkb1* knockout mice have lower serum IgG1 and IgE antibodies [28]. Stat6 induces *Ighg1* germline transcription, an event that is obligatory for the intrachromosomal DNA recombination that leads to IgG1 switching [29]. Notably, we observed an increase in *Bach2* expression during the middle transcriptional phase (Fig 4D). *Bach2* is required for both CSR and SHM and regulates *Aicda* expression [30,31]. We next analyzed expression profiles of cytokine receptors, signaling by which is known to influence CSR. External signaling from IL-4 is known to drive IgG1 switching [32]. We observed higher expression of *Il4ra*, encoding an IL-4 receptor subunit (Fig 4F) within the early transcriptional phase. In contrast, *Ifngr1* transcript expression was low during the early phase consistent with the findings that IFN-γ signaling inhibits IgG1 switching [33]. These results suggest a

regulatory model where such B cells are poised to receive IL-4 signaling and undergo IgG1 switching. In conclusion, pseudocell tracing inferred expression dynamics of *Aicda*, *Il4ra*, *Ifngr1*, *Stat6 and Bach2* in B cells undergoing CSR to the IgG1 isotype, that were consistent with prior biological knowledge and also generated new hypotheses.

Discussion

We present a supervised machine learning framework, Pseudocell Tracer, for modeling cellular trajectories in complex systems. Inference of cellular trajectories from scRNAseq datasets remains a challenging problem, particularly for heterogeneous cellular compartments with complex dynamics. Existing methods for trajectory inference are strongly dependent on initial low-dimensional projections of the datasets. Concurrent and inter-digitated transcriptional programs, such as those regulating cell cycle and metabolism can obstruct unsupervised dimensionality reduction and pseudotemporal ordering techniques. This problem can be amplified by sparse sampling of key intermediate cellular states. To address these two challenges, we harness adjacent biological information in order to shape the latent space in a biologically valid manner thereby revealing discrete cellular trajectories in a complex developmental compartment that are otherwise obscured.

By formally utilizing adjacent biological information Pseudocell Tracer seeks to complement the existing tools for trajectory inference, which analyze scRNAseq datasets without initial reference to biological knowledge of the system. Pseudocell Tracer learns a generative model encompassing all cells in a biologically meaningful latent space. As a result, the generative model provides a means to interpolate cells in the latent space and allows for the specific delineation of pseudocells by conditioning on adjacent biological information. Importantly, we demonstrate that even a relatively small dataset with a few hundred cells is sufficient for learning and can be used to generate biologically plausible virtual cells. The surprising effectiveness of GANs in simulating realistic subpopulations of cells from small datasets has been independently demonstrated [20].

Pseudocell Tracer was used to analyze and infer gene expression dynamics along a particular CSR trajectory (IgG1), during a prototypic antigen-induced B cell response. In spite of extensive genetic and molecular analysis of CSR, the gene expression dynamics of B cells undergoing CSR in vivo have not been revealed. In fact, a recent report using extensive scRNAseq profiling of human tonsillar B cells was still unable to reveal the developmental modulation of genomic states underlying particular CSR trajectories using unsupervised dimensionality reduction techniques [34]. We utilized a unique scRNAseq dataset generated from antigen-specific B cells induced by NP-KLH immunization to infer the cellular trajectories of germinal center B cells undergoing CSR to IgG1, the dominant isotype manifested under these conditions. Although recent work has indicated that CSR is primarily induced before entry of antigen-specific B cells into germinal centers [35], we were able to detect it also within the GC compartment.

Our results revealed an ordering of key transcription factors regulating CSR, including the higher expression of *Nfkb1* and *Stat6* prior to the upregulation of *Bach2* expression. The former transcription factors have known roles in regulating IgG1 CSR and the latter promotes *Aicda* gene expression and therefore both CSR and SHM. An intriguing finding is the upregulation of *Aicda* gene expression in the CSR trajectory after it has been initiated suggesting that increased expression of AID maybe needed for efficient SHM which occurs in B cells within the germinal centers. Finally, the expression pattern of genes encoding cytokine receptors within the early transcriptional phase revealed a poised B cell state that maybe committed to CSR to the IgG1 isotype.

Ultimately, the hypotheses generated through Pseudocell Tracer will have to be experimentally validated. While we demonstrate proof-of-concept in modeling CSR along a particular isotype trajectory, we anticipate future studies analyzing other isotypes and facilitate assembly of isotype-specific B cell trajectories in diverse lymphoid organs and tissues. In the current work, we utilized an encoding that reflected potential CSR paths, however, Pseudocell Tracer can encode other structured adjacent biological information as well, such as phylogenetic trees constituted by somatically mutating antibody variable regions. In so doing, Pseudocell Tracer could be used to guide the latent space and conditional generation of specific trajectories of B cells undergoing SHM and affinity maturation.

The machine learning framework underlying Pseudocell Tracer provides for a flexible means to explore new forms of adjacent biological information in extremely diverse contexts. Ultimately, Pseudocell Tracer is a powerful framework for characterizing the transcriptional states and trajectories of cells during their development and activation. These states and trajectories, particularly rare ones, can be revealed by embedding them in valid biological priors. For example, Pseudocell Tracer has the potential to merge multiple single cell profiling experiments from different biological compartments, providing a novel way to bridge datasets by using prior molecular information about their relatedness. Therefore, Pseudocell Tracer promises to be a robust engine for hypothesis generation for experimental biology by predicting novel regulators and rare cell states underlying extremely diverse cellular trajectories.

Materials and methods

Ethics statement

All mice used for experiments were maintained under specific-pathogen-free conditions and the immunization experiments were approved by the University of Pittsburgh Institutional Animal Care and Use Committee under protocol no. 19115454.

Mice and immunization

C56BL/6J (Jax 000664) mice were obtained from the Jackson Laboratory. Mice were housed in specific pathogen-free conditions and were used and maintained in accordance of University of Pittsburgh Institutional Animal Care and Use Committee guidelines. Six to eight week old mice were immunized intraperitoneally with 100 μg NP(23)-KLH (Biosearch Technologies) mixed with 50% (v/v) Alum (Thermo Scientific) and 1 μg LPS (Sigma).

Sorting of NP-specific B cells and scRNAseq

Mouse spleens were collected on day 14 post-NP-KLH immunization. Splenocytes were washed and prepared as single-cell suspensions in MACS buffer (pH 7.4; PBS without calcium and magnesium plus 3% FBS and 2 mM EDTA). Cells were blocked with 25 μg/ml 2.4G2 (BD) for 15 minutes on ice and labeled with viability dye eFluor 780 (eF780), 0.04 μg/mL NP_{138}-PE and 2 μg/mL B220-APC (RA3-6B2) for 30 minutes at 4˚C. B cells were sorted as $eF780^{-}B220^{+}NP^{+}$ using FACSAria II (BD) with 70 μm nozzle at 4˚C. FACS sorted $B220^{+}NP^{+}$ B cells were mixed with reverse transcription reagents and loaded onto the Chromium instrument (10x Genomics) and libraries were prepared. Single cell libraries were sequenced using Illumina's NovaSeq platform.

scRNAseq data preprocessing

Genes containing no counts across any samples were discarded. We calculated the relative isotype expression within a B cell by dividing the log2(TPM + 1) expression of each distinct isotype transcript with the cumulative expression of all isotypes. The relative isotype expression

profiles were clustered using hierarchical clustering using the *clustergram* function in Matlab with default settings.

Pseudotime trajectory

We apply two state-of-the-art pseudotime trajectory inference methods to our data: Monocle3, and Slingshot. Monocle3 is a method to learn pseudotime through the use of dimension reduction and graph learning. A minimum spanning tree is constructed and used to order the cells to infer pseudotime trajectory. We implemented Monocle3 using the default settings with UMAP dimension reduction and Louvain clustering. Slingshot is another single cell trajectory inference method. In their method, they project the data into a latent space and perform clustering. In our implementation of Slingshot, we use PCA for dimension reduction and Gaussian mixture model (GMM) clustering with default parameters. The minimum spanning tree is then constructed from the clusters.

Supervised autoencoder

In order to shape the latent space in a meaningful way, we utilize a supervised autoencoder to help distinguish important differences in cell subtypes. Our work constructs a supervised autoencoder in two steps. The first step employs supervised encoding using the relative abundance of the IgH genes of interest based on their relative expression. To do so, the normalized gene counts are encoded into a latent layer using a neural network model, and this latent layer is then used to predict the relative abundance of the different IgH genes. The encoder has an input of size 6,617 and contains two fully connected layers between the input and the latent layer of sizes 512 and 256 respectively, each using the rectified linear unit (ReLU) activation function. The latent layer has a size of 64 nodes and uses the sigmoid activation function. In addition, there is a fully connected layer of size 128 between the latent layer and the output. The output layer contains 8 nodes and uses the softmax activation function in order to generate a relative distribution across the 8 IgH genes. The network is trained using the Adam optimizer with a learning rate of $1x10^{-4}$ and the Kullback-Leibler divergence (KLD) loss function,

$$L_{enc} = \frac{1}{n}\sum_{i}^{n} \mathrm{KLD}(Y_i||\hat{Y}_i) + \lambda \sum_{l} |W_l|^2$$

Here, the first term represents the average KLD between the observed IgH proportions, Y_i, and the predicted IgH proportions, $\hat{Y}_i$, across all n samples. The second term regularizes the network using the weighted sum the L2-norms for the weights of each layer W_l, *where* l represents the layer. In our study, we found the best weight coefficient to be $\lambda = 1x10^{-5}$. Further regularization is added by implementing dropout with a rate of 0.3 at every hidden layer. The network is trained using early stopping, where 10% of the training data is held out as a validation set. The KLD term of the loss function was evaluated on the validation set each epoch and training was terminated if there had not been a decrease in 100 epochs, whereafter the model was reverted to the previously best state. The model is then trained for an additional 5 epochs using the entire training data, including the validation set that was used for early stopping.

After the supervised encoder had been trained, the second step employed a decoder in order to reconstruct the original normalized gene values from the latent space. This network takes the latent space of size 64 as an input to predict the normalized gene expression for each of gene as an output. The network contains two fully connected layers of sizes 256 and 512 respectively, each using the ReLU activation function. The output layer has a size of 6,617 and uses the linear activation function. The network is trained using the Adam optimizer with a

learning rate of $1x10^{-4}$ with the mean squared error (MSE) loss function,

$$L_{dec} = \frac{1}{n}\sum_{i}^{n}(X_i - \hat{X}_i)^2 + \lambda \sum_{l} |W_l|^2$$

Here, the first term represents the average MSE between the observed normalized gene expression, X_i, used to generate the latent space of some sample i, and the decoded normalized gene expression $\hat{X}_i$. The second term regularizes the network using the weighted sum the L2-norms in the same way as seen in the supervised encoder. We again found the best regularization parameter to be $\lambda = 1x10^{-5}$. The decoder model was further regularized using batch normalization at each of the hidden layers. Parameter choices are listed in S2 Table. The model was trained using the same early stopping approach from the supervised encoder, however in this case we stop the training based on the MSE term of the loss function.

Conditional generative adversarial network

In order to generate samples for unobserved pseudo-times, we utilize a CGAN architecture [36]. A CGAN is composed of two networks: a generator and a discriminator. The generator model learns to generate fake data that is as close to the distribution of the real data as possible. At the same time, the discriminator model tries to predict if a piece of data is real or fake. Both models are trained in an adversarial manner where the generator tries to maximize the log-probability of labeling real and fake images correctly while the discriminator tries to minimize it, resulting in a zero-sum minimax game. The models are trained using gradient descent until Nash equilibrium is reached.

The generator in our model, G, takes in as an input a vector of 32 values sampled from ~U(−1,1) concatenated with the relative abundance of the 8 IgH genes in order to output a vector representing the encoded latent representation of gene expression values. The network contains two fully connected layers of sizes of 256 and 512 respectively, both using the ReLU activation function. The output is size 64 with a linear activation function. The discriminator, D, takes a vector of size 64 representing encoded gene expression as well as the relative abundance of the 8 IgH genes and outputs a single value between 0 and 1 as the probability of the data being real. The discriminator has a single layer of size 512 that uses the ReLU activation function. To avoid problems of a vanishing gradient, we adjust the loss function of the generator to create a non-saturating loss function [36]. The networks are trained simultaneously using the Adam optimizer with a learning rate of $5x10^{-5}$ until convergence (about 50,000 epochs). The loss function for the discriminator and generator are shown below.

$$L_D = \frac{1}{n}\sum_{i}^{n} -\log[D(z_i)] - \log[1 - D(G(r_i))]$$

$$L_G = \frac{1}{n}\sum_{i}^{n} \log[D(G(r_i))]$$

Here z_i represents the latent space generated from the observed gene expression from sample i using the and $G(r_i)$ represents the generated latent space using the relative IgH values from sample i. Latent spaces are taken from the autoencoder model trained on the entire data set.

Pseudocell generation

Once the CGAN model was trained to equilibrium, we generated 100 latent spaces for each relative isotype expression profile. For each of IgG1, IgG2b, and IgG3, we generated latent spaces for each trajectory starting at 100% IgM, and for each subsequent relative isotype expression profile, we reduced the relative abundance of IgM by 1% and increased the respective isotype gene expression by 1%. We then decoded each latent space back to the normalized gene expression values in order to obtain pseudo cells. Gene expression trajectories was plotted using the average and 95% confidence interval along the trajectory. Concordance between the nearest real cell to each pseudo cell was qualitatively evaluated for *Ighm* and *Aicda* (S5 Fig) to evaluate model behavior.

Supporting information

S1 Fig. UMAP of RNA-seq data, colored by cell cycle genes.
(EPS)

S2 Fig. Detailed architecture of the supervised autoencoder.
(EPS)

S3 Fig. Detailed architecture of the conditional GAN.
(EPS)

S4 Fig. Plot of relative expression of *Aicda*, *Ighm* and *Ighg2b* (left) *and Ighg3* (right).
(EPS)

S5 Fig. Predicted and observed target gene expression along IgM–IgG1 trajectory.
(EPS)

S1 Table. Sample size analysis.
(DOCX)

S2 Table. Hyperparameter table.
(DOCX)

Acknowledgments

The authors thank Nicolas Chevrier, Sasha Chervonsky, and members of the Khan lab for helpful discussions about this work.

Author Contributions

Conceptualization: Heping Xu, Harinder Singh, Aly A. Khan.

Formal analysis: Derek Reiman, Nathan Salomonis.

Methodology: Derek Reiman, Godhev Kumar Manakkat Vijay, Andrew Sonin, Dianyu Chen, Aly A. Khan.

Resources: Heping Xu, Harinder Singh, Aly A. Khan.

Software: Derek Reiman.

Supervision: Harinder Singh, Aly A. Khan.

Validation: Godhev Kumar Manakkat Vijay, Heping Xu.

Writing – original draft: Harinder Singh, Aly A. Khan.

Writing – review & editing: Harinder Singh, Aly A. Khan.

References

1. Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium. Nature. 2018; 562 (7727):367. https://doi.org/10.1038/s41586-018-0590-4 PMID: 30283141
2. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. Science forum: the human cell atlas. Elife. 2017; 6:e27041. https://doi.org/10.7554/eLife.27041 PMID: 29206104
3. Neu KE, Tang Q, Wilson PC, Khan AA. Single-cell genomics: approaches and utility in immunology. Trends in immunology. 2017; 38(2):140–9. https://doi.org/10.1016/j.it.2016.12.001 PMID: 28094102
4. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nature biotechnology. 2019; 37(5):547–54. https://doi.org/10.1038/s41587-019-0071-9 PMID: 30936559
5. Kakaradov B, Arsenio J, Widjaja CE, He Z, Aigner S, Metz PJ, et al. Early transcriptional and epigenetic regulation of CD8+ T cell differentiation revealed by single-cell RNA sequencing. Nature immunology. 2017; 18(4):422. https://doi.org/10.1038/ni.3688 PMID: 28218746
6. Yu Y, Tsang JC, Wang C, Clare S, Wang J, Chen X, et al. Single-cell RNA-seq identifies a PD-1 hi ILC progenitor and defines its development pathway. Nature. 2016; 539(7627):102. https://doi.org/10.1038/nature20105 PMID: 27749818
7. Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H, et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature. 2016; 537(7622):698. https://doi.org/10.1038/nature19348 PMID: 27580035
8. Neu KE, Guthmiller JJ, Huang M, La J, Vieira MC, Kim K, et al. Spec-seq unveils transcriptional subpopulations of antibody-secreting cells following influenza vaccination. The Journal of Clinical Investigation. 2019; 129(1):93–105. https://doi.org/10.1172/JCI121341 PMID: 30457979
9. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016; 167(7):1853–66. e17. https://doi.org/10.1016/j.cell.2016.11.038 PMID: 27984732
10. Bossel NB-M, Hen-Avivi S, Levitin N, Yehezkel D, Oosting M, Netea M, et al. Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. Nature communications. 2019; 10(1):3266–. https://doi.org/10.1038/s41467-019-11257-y PMID: 31332193
11. Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, et al. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. Science immunology. 2017; 2(9). https://doi.org/10.1126/sciimmunol.aal2192 PMID: 28345074
12. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell. 2017; 169(7):1342–56. e16. https://doi.org/10.1016/j.cell.2017.05.035 PMID: 28622514
13. Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nature immunology. 2019; 20(7):928. https://doi.org/10.1038/s41590-019-0378-1 PMID: 31061532
14. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single Cell RNAseq Analysis. bioRxiv. 2019:641142.
15. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics. 2018; 19(1):477. https://doi.org/10.1186/s12864-018-4772-0 PMID: 29914354
16. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology. 2014; 32(4):381. https://doi.org/10.1038/nbt.2859 PMID: 24658644
17. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome biology. 2019; 20(1):59. https://doi.org/10.1186/s13059-019-1663-x PMID: 30890159
18. Herring CA, Banerjee A, McKinley ET, Simmons AJ, Ping J, Roland JT, et al. Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. Cell systems. 2018; 6(1):37–51. e9. https://doi.org/10.1016/j.cels.2017.10.012 PMID: 29153838
19. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology. 2019; 15(6):e8746. https://doi.org/10.15252/msb.20188746 PMID: 31217225

20. Marouf M, Machart P, Bansal V, Kilian C, Magruder DS, Krebs CF, et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. Nature Communications. 2020; 11(1):1–12. https://doi.org/10.1038/s41467-019-13993-7 PMID: 31911652

21. Ghahramani A, Watt FM, Luscombe NM. Generative adversarial networks simulate gene expression and predict perturbations in single cells. BioRxiv. 2018:262501.

22. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications. 2019; 10(1):1–14. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

23. Stavnezer J, Guikema JE, Schrader CE. Mechanism and regulation of class switch recombination. Annu Rev Immunol. 2008; 26:261–92. https://doi.org/10.1146/annurev.immunol.26.021607.090248 PMID: 18370922

24. Manis JP, Tian M, Alt FW. Mechanism and control of class-switch recombination. Trends in immunology. 2002; 23(1):31–9. https://doi.org/10.1016/s1471-4906(01)02111-1 PMID: 11801452

25. Jacob J, Kassir R, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl. I. The architecture and dynamics of responding cell populations. The Journal of experimental medicine. 1991; 173(5):1165–75. https://doi.org/10.1084/jem.173.5.1165 PMID: 1902502

26. Furukawa K, Akasako-Furukawa A, Shirai H, Nakamura H, Azuma T. Junctional amino acids determine the maturation pathway of an antibody. Immunity. 1999; 11(3):329–38. https://doi.org/10.1016/s1074-7613(00)80108-9 PMID: 10514011

27. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nature methods. 2017; 14(10):979. https://doi.org/10.1038/nmeth.4402 PMID: 28825705

28. William CS, Liou H-C, Tuomanen EI, Baltimore D. Targeted disruption of the p50 subunit of NF-κB leads to multifocal defects in immune responses. Cell. 1995; 80(2):321–30. https://doi.org/10.1016/0092-8674(95)90415-8 PMID: 7834752

29. Harris MB, Chang C-C, Berton MT, Danial NN, Zhang J, Kuehner D, et al. Transcriptional repression of Stat6-dependent interleukin-4-induced genes by BCL-6: specific regulation of Iε transcription and immunoglobulin E switching. Molecular and cellular biology. 1999; 19(10):7264–75. https://doi.org/10.1128/mcb.19.10.7264 PMID: 10490661

30. Igarashi K, Ochiai K, Itoh-Nakadai A, Muto A. Orchestration of plasma cell differentiation by Bach2 and its gene regulatory network. Immunological reviews. 2014; 261(1):116–25. https://doi.org/10.1111/imr.12201 PMID: 25123280

31. Budzyńska PM, Kyläniemi MK, Kallonen T, Soikkeli AI, Nera KP, Lassila O, et al. Bach2 regulates AID-mediated immunoglobulin gene conversion and somatic hypermutation in DT40 B cells. European journal of immunology. 2017; 47(6):993–1001. https://doi.org/10.1002/eji.201646895 PMID: 28301039

32. Higgins BW, McHeyzer-Williams LJ, McHeyzer-Williams MG. Programming isotype-specific plasma cell function. Trends in immunology. 2019. https://doi.org/10.1016/j.it.2019.01.012 PMID: 30846256

33. Kawano Y, Noma T, Yata J. Regulation of human IgG subclass production by cytokines. IFN-gamma and IL-6 act antagonistically in the induction of human IgG1 but additively in the induction of IgG2. The Journal of Immunology. 1994; 153(11):4948–58. PMID: 7963558

34. King HW, Orban N, Riches JC, Clear AJ, Warnes G, Teichmann SA, et al. Antibody repertoire and gene expression dynamics of diverse human B cell states during affinity maturation. bioRxiv. 2020:2020.04.28.054775. https://doi.org/10.1101/2020.04.28.054775

35. Roco JA, Mesin L, Binder SC, Nefzger C, Gonzalez-Figueroa P, Canete PF, et al. Class-switch recombination occurs infrequently in germinal centers. Immunity. 2019; 51(2):337–50. e7. https://doi.org/10.1016/j.immuni.2019.07.001 PMID: 31375460

36. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al., editors. Generative adversarial nets. Advances in neural information processing systems; 2014.