VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics

Lucas Seninge¹, Ioannis Anastopoulos¹, Hongxu Ding¹ & Joshua Stuart¹

Deep learning architectures such as variational autoencoders have revolutionized the analysis of transcriptomics data. However, the latent space of these variational autoencoders offers little to no interpretability. To provide further biological insights, we introduce a novel sparse Variational Autoencoder architecture, VEGA (VAE Enhanced by Gene Annotations), whose decoder wiring mirrors user-provided gene modules, providing direct interpretability to the latent variables. We demonstrate the performance of VEGA in diverse biological contexts using pathways, gene regulatory networks and cell type identities as the gene modules that define its latent space. VEGA successfully recapitulates the mechanism of cellular-specific response to treatments, the status of master regulators as well as jointly revealing the cell type and cellular state identity in developing cells. We envision the approach could serve as an explanatory biological model for development and drug treatment experiments.
Recent advances in single-cell RNA sequencing (scRNA-Seq) technologies have enabled the characterization of cellular states at an unprecedented scale and resolution\(^1\). Among the many widely-used frameworks for analyzing complex transcriptomic patterns in single cells, artificial neural networks (ANNs) such as autoencoders (AEs)\(^2\) have emerged as powerful tools. AEs are neural networks that transform an input dataset into a decoded representation while minimizing the information loss\(^3\). The diversity in their architectural design makes AEs suitable to tackle various important challenges of scRNA-Seq such as dimensionality reduction\(^4\), clustering\(^5\), and data denoising\(^6\).

More recently, deep generative models such as variational autoencoders\(^7\) (VAEs) have proven to be extremely useful for the probabilistic modeling of single-cell transcriptomes, such as scVI and scGen\(^8\)–\(^10\). While standard AEs learn to reconstruct an input dataset, deep generative architectures explicitly model and learn the true data distribution, which allows a broader set of queries to be addressed. While deep generative models have shown impressive performance for their dedicated modeling tasks, they often lack interpretability thus cannot offer a biologically meaningful latent representation of transcriptomes. For example, latent perturbation vectors extracted with scGen cannot be directly related to gene module variations\(^10\).

Integration of prior knowledge about gene modules to aid interpretability has already been successfully applied to transcriptomics data. DCell\(^11\) is a deep neural network integrating the hierarchical information about the molecular subsystems involved in cellular processes to guide supervised learning tasks, such as predicting growth in yeast. Such a model yields an informative biological interpretation of predictions by investigating the activation of the different subsystems embedded in the model's architecture. However, this model only works in a supervised learning setting where the goal is to predict a phenotypic outcome. On the other hand, f-scLVM\(^12\) is a Bayesian hierarchical model with explicit prior biological knowledge specification to infer the activity of latent factors as a priori characterized gene modules. While this approach enables the modeling of single-cell transcriptomes in an interpretable manner, the computational cost of the inference algorithm, as well as the absence of inference for out-of-sample data, make the development of more efficient approaches highly desirable.

Here we propose VEGA (VAE enhanced by gene annotations), a VAE with a sparse linear decoder informed by biological networks. VEGA offers an interpretable latent space to represent various biological information, e.g., the status of biological pathways or the activity of transcriptional regulators. Specifically, the scope of VEGA is twofold, (1) encoding data over an interpretable latent space and (2) inferring gene module activities for out-of-sample data.

**Results**

**Architectural design of VEGA.** To create a readily interpretable VAE, we propose a novel architecture we refer to as VEGA (VAE enhanced by gene annotations) where the decoder (generative part) connections of the neural network are guided by gene module membership as recorded in gene annotation databases (e.g., Gene Ontology, PANTHER, MolSigDB, or Reactome) (Fig. 1a). In many standard VAE implementations, the information bottleneck of the encoder-decoder architecture often represents latent variables modeled as a multivariate normal distribution. Despite providing highly informative representations of the input data, VAE latent variables are in general hard to interpret. Svensson et al.\(^13\) proposed using a linear decoder which directly connects latent variables to genes, providing interpretability similar to that offered by standard factor models such as PCA. Although providing valuable insights, such an approach requires further statistical enrichment tests on the weights of the decoder to infer biological processes contributing to the single-cell expression dataset.

In contrast to previous approaches, VEGA implements a sparse architecture that explicitly reflects knowledge about gene regulation. In the service of biological pathways, genes work together in gene modules, regulated by common transcription factors that often produce correlated expression. Thus, if a given scRNA-Seq dataset \(X\) reflects the patterns of known gene modules, then it is possible for a VAE to learn a compact representation of the data by incorporating those modules as latent variables \(Z\). VAEs use multiple layers to approximate the latent variable distribution and produce a low dimensional, nonlinear representation of the original feature data. Importantly, the first and last layers directly connect to the input or predicted features and so can be fashioned to depict intuitive groupings. Standard VAEs use a fully connected layer for both the encoding first layer and the decoding final layer (SFig. 1a). Instead, VEGA uses a gene membership mask \(M\) to select a subset of trainable weights in the decoder layer that are determined by a given set of gene modules (see Methods). The mask is applied to the weights that connect to the predicted output features to yield an interpretation of the latent variable layer where each latent variable is viewed as a specific gene module, henceforth referred to as a gene module variable (GMV). Specifically, the generative part of VEGA (decoder) maintains a link from a GMV to an output gene only if this gene is annotated to be a member of this specific gene module. The two main advantages of this design are (1) the latent variables are directly interpretable as the activity of biological modules and (2) the flexibility in the gene module specification allows it to generalize to different biological abstractions (such as pathways, gene regulatory networks (GRNs), or even cell types) and can be taken from any of several curated databases of gene sets (such as MSigDB\(^14\), Reactome pathways\(^15\), inferred GRNs\(^16\)). Additionally, VEGA incorporates information about covariates such as technical replicates in its latent space. This can be used to alleviate batch effects, as it has been demonstrated in previous deep generative models for single-cell data\(^9\) (Fig. 1a and SFig. 2).

Note that it is possible to implement gene module sparseness in the encoder half of the neural network (inference part), in addition to (or in place of) the decoder half (generative part), which gives three possible VAE architectures that we considered for single-cell RNA-seq analysis (SFig. 1ai–iii). As expected, we found that the GMV-guided designs resulted in decent although slightly worse performance compared to the full architecture (SFig. 1c). Among these options, we chose the sparse decoding architecture over the others for its improved separation of known cellular states and types in the Kang et al. PBMC data\(^17\) (SFig. 1b). Intuitively, using a deep encoder maintains a full VAE’s inference capacity to capture a potentially complex latent space while together with a sparse decoder approximates the posterior distribution of GMV activities \(p(Z|X)\) to provide interpretation over gene modules. Additionally, we found that VEGA benefits from having a trainable, sparse decoder to adequately capture the biological signal of a dataset compared to simpler pathway transformations (SFig. 3).

Recapitulating biological information over an interpretable latent space. We asked if VEGA could recapitulate the status of biological pathways by applying it to a published and well-studied peripheral blood mononuclear cells (PBMCs) dataset stimulated with the chemokine interferon-\(β\)\(^7\) (Methods). We first found that VEGA is able to capture cell types and stimulation status using the Reactome collection of processes and pathways\(^15\) in the GMV decoding layer.
Fig. 1 Designing a novel VAE architecture with interpretable latent space. a Overview of the VEGA model. Composed of a deep nonlinear encoder ($\mu, \Sigma$) and a masked linear decoder, VEGA represents single-cell transcriptomics data into a lower-dimensional interpretable latent space $z$ that approximates a set of user-supplied gene modules (GMV). Additionally, VEGA can integrate batch information as another variable $s$ to condition its generative process on batch labels. b UMAP embedding of the latent space of VEGA retains the biological signal of the Kang et al. PBMCs dataset17. c Inferred interferon-alpha/ beta signaling pathway activity segregates stimulated cells from the control population. d Bivariate GMV plot showing the ability of the model to recover the tryptophan catabolism activity, an innate (Dendritic cells, FCGRTA+monocytes, CD14+monocytes, CD14+monocytes) immune cell-specific response to the perturbation. e Volcano plot showing differentially active GMVs between stimulated and control innate immune cells. The red dots indicate GMVs with $\log_{10}$(Bayes Factor)$ > 3$ and a mean absolute difference (MD) in the latent space of at least 5. f Comparison of VEGA Bayes Factor with GSEA $-\log_{10}$(FDR). The size of the dots indicates the gene set size. The red, blue, and purple quadrants correspond respectively to significant hits unique to our model, unique to GSEA, and common to both.

(Fig. 1b). Specifically, we found that the interferon-$\alpha/\beta$ signaling GMV activity segregates stimulated and naive cells, confirming the ability of VEGA to capture pathway activity in its latent space (Fig. 1c, d). We further examined other known biological pathways involved in interferon-induced immune cell activation and found cell-type-specific activation of certain cellular processes. For example, tryptophan catabolism response to interferon separates innate immune cells (Dendritic cells, FCGRTA+monocytes, and CD14+monocytes) from adaptive immune cells (NK cells, T-cell CD8, T-cell CD4, and B cells) (Fig. 1d), as previously investigated18,19. Together, these results suggest that VEGA’s GMV’s reflect the expected major biological pathways in PBMCs and therefore may be useful for other datasets to project cells into an interpretable space, allowing investigation of cell-type-specific patterns at the cellular process level.

We next asked whether the differential activities of the GMVs accurately contrast pathway states as a function of a specific, experimentally controlled context.

For this purpose, we propose a similar Bayesian hypothesis testing procedure as introduced by Lopez et al.9 to study the difference in GMV activities. As VEGA models the posterior distribution of each GMV, we can formulate mutually exclusive hypotheses similar to differential gene expression tests (i.e., GMVs are activated at different levels). We can approximate the posterior probability of these hypotheses through Monte Carlo sampling of VEGA’s latent variable distribution. The ratio of hypothesis probabilities corresponds to the Bayes Factor20 (BF, see Methods).

When applied to innate immune cells in the stimulated vs control groups of the Kang et al.17 dataset, the BF analysis found GMVs that correspond to pathways expected to be activated in the stimulated groups (interferon signaling, tryptophan catabolism; $\log_{10}$(BF)$ > 3$, Fig. 1e). We compared the GMV BFs with the false discovery rate (FDR) values of the standard GSEA toolkit (Methods, Fig. 1f). While both methods found the expected activation of the interferon-$\alpha/\beta$ signaling pathway GMV in the stimulated groups, GSEA missed the tryptophan catabolism activation in innate immune cells (Fig. 1f). Overall, VEGA seems more robust than GSEA to gene set size bias (Fig. 1f and Sfig. 4), suggesting it may emphasize more context-relevant pathways. Additionally, the differential GMV activity test can be applied in a cell-type-specific fashion (similar to one-vs-rest differential gene expression analyses). We found that such a procedure yields informative results in terms of cell type-specific biological processes activated independently of perturbation status (Sfig. 5 and Supplementary Data 1).

Large-scale investigation of biological responses to drug treatments in cell lines. Next, we investigated whether VEGA could detect patterns of drug responses in large-scale experiments over cancer cell lines, such as the data introduced in recent experimental protocols like MIX-Seq21. To this end, we gathered single-cell data for 97 cancer cell lines under five different conditions: 24 h DMSO treatment (control), 24 h Trametinib treatment (MEK inhibitor), 24 h Dabrafenib treatment (Mutated BRAF inhibitor), 24 h Navitoclax treatment (Bcl-2 inhibitor), and 24 h BRD3379 treatment (tool compound with unknown mode of action, MoA) (Methods). We trained one model for each different drug treatment (four models in total) by combining the drug treatment dataset and the control group (DMSO dataset), initializing the GMVs of VEGA with the hallmark gene sets from MSigDB22 to focus on core cellular processes. Overall, each model was able to separate cell lines and treatment conditions in the GMV space (Fig. 2a, and Sfig. 6). For Trametinib notably, the important change in G2M checkpoint GMV activity (decrease in the treated condition) agrees with the expected MoA of a MEK inhibitor23,24 (Fig. 2b). Next, we sought to investigate whether we could recapitulate the pattern of biological responses between control and treated conditions for each cell line/drug treatment pair. For each pair, we computed GMV BFs to approximate differential pathway activities between the two conditions. The resulting heatmap can be used to understand and interpret patterns of response over all experimental conditions (Fig. 2c). As found when visually investigating the low dimensional...
Latent space (Fig. 2d). We examined the activity of STAT3 and found that the pre-annotated cell types were well-separated in the predicted target genes of each transcription factor. After training, we created a GMVs decoding architecture, which was set to the reported GMVs by the original authors (Fig. 3b). Moreover, in a one-vs-rest differential analysis setting for each cell type population, the activity of the GMVs was largely anticorrelated in neoplastic cells (Fig. 3c). We also found that Dabrafenib-treated BRAF-mutant melanoma cell lines exhibited larger loge(BF) than Dabrafenib-treated cell lines (average loge(BF)) of 0.763 vs 0.668 for other cell lines), clustering with the Trametinib-treated cell lines as reported in the MIX-Seq study (Fig. 2c and SFig. 6d). Overall, the results presented here agree with the previous gene set analysis results on this dataset, and demonstrate VEGA’s GMVs can recapitulate patterns of drug response in large-scale experiments.

Gene regulatory analysis of glioblastoma reveals stratification of neoplastic cells. As mentioned previously, one of VEGA’s strengths is the flexibility in the specification of the GMV connectivity, as any gene module can be used in the decoder. Transcription factors often exert tight regulation of gene expression programs and phenotypes. To this end, we investigated whether using master transcriptional regulators as the GMVs could help understand the underlying GRNs in the context of a single-cell glioblastoma (GBM) dataset. We used the GBM ARACNe network reported in Carro et al. to guide the structural design of our model. Specifically, VEGA’s GMVs were set to the reported transcription factors and the connectivity matrix M, defining the GMVs decoding architecture, was created from the set of predicted target genes of each transcription factor. After training, we found that the pre-annotated cell types were well-separated in the latent space (Fig. 2d). We examined the activity of STAT3 and OLIG2, two well-known master regulators of the mesenchymal (MES) and proneural (PN) GBM subtypes, respectively. We confirmed that their GMV activity was largely anticorrelated in neoplastic cells (Fig. 2e). Additionally, OLIG2, a known master regulator of oligodendrocyte differentiation, was inferred as activated in oligodendrocyte precursor cells (OPCs). These results demonstrate that VEGA is able to home-in on the relevant transcriptional regulators when the decoder wiring is extended to model known factor-to-target relationships.

Combining cell type and cellular state representations refines cortical organoid development analysis. A great challenge of modern cellular biology is to identify and define cell types and cellular states, at the level of individual cells, in order to systematically study homeostasis and disease development under a common vocabulary. In a typical single-cell study, a few “marker sets” will be known, each containing a list of genes having expected expression patterns for some of the cell types of interest. Leveraging such marker sets often provides clues and helps orient data analysis. We asked whether the information recorded in such marker sets could be used in VEGA to produce a disentangled representation of cell types and cellular states. To this end, we added a GMV zt with appropriate entries in M, for each latent cell type t in addition to the Reactome pathway GMVs already in VEGA’s model.

We applied VEGA to a dataset of cells assayed during the early development of cortical organoids from Field et al., including all of the major cell types defined in the study as GMVs (Fig. 3a). After training, we found that the activity of each marker set GMV was able to correctly segregate its corresponding cell type as annotated by the original authors (Fig. 3b–d). Moreover, in a one-vs-rest differential GMV analysis setting for each cell type population, the activity of the corresponding marker set GMV showed significant enrichment ([loge(BF)] > 3), which suggests using GMV BFs could help annotate the cell types of unknown clusters (Fig. 3e). We further noted that the...
most differentially activated GMVs were coherent in the context of early brain development (Fig. 7 and Supplementary Data 2). To study whether VEGA could separate cell type identity from cellular states such as dividing vs quiescent cell populations, we projected the dataset into two components: (1) the cell type GMV representing the neural epithelium marker set (a type of early brain progenitor) and (2) the cell state GMV representing the cell cycle mitotic pathway activity (Fig. 3f). As discussed previously, the activity of the neural epithelium GMV separated the neural epithelium cells from the rest of the dataset, while the activity of the cell cycle mitotic pathway GMV separated quiescent from actively dividing cells in the two progenitors populations (radial glia cells and neural epithelium). To validate that the cells identified as dividing were proliferating, we studied the correlation between the cell cycle mitotic pathway GMV activity and the expression of the MKI67 gene, a canonical marker of proliferating (external validator not present in the cell cycle mitotic pathway set) (Fig. 3g). Overall, the expression of MKI67 correlates well with the inferred activity of the cell cycle mitotic pathway GMV ($R^2 = 0.64$). Together, these results demonstrate VEGA’s potential use to jointly infer cell type and state for different populations of cells, as combining different sources of information (pathways, master regulators, and cell type markers) in the latent space can shed light on different aspects of the identity of a single-cell.

**Generalization of the inference process to out-of-sample data.**

We next asked whether VEGA could generalize to correctly infer an interpretable latent representation of data unseen at the time of training (out-of-sample data). To this end, we evaluated VEGA in two settings. In the first case, we measured the biological generalization of VEGA’s inference by holding out (cell type, condition) pairs during training. Specifically, we investigated whether the inferred GMV activities for held-out cells were conveying the same biological information as to when this population is seen at the time of training. To this end, we removed one cell type of the stimulated condition during training, and then inferred the GMV activities for that held-out population (out-of-sample) and compared them to the GMV activities learned from the fully trained model. The experiment was conducted using the Kang et al. PBMC dataset. In the second case, we estimated the “technical generalization” of VEGA’s inference by training on one dataset (study A) and then evaluating on a second dataset (study B) that contains only control cells. We used the Kang et al. PBMC dataset as study A and the Zheng et al. dataset as study B.

For the biological generalization test, we first checked that the distribution of the interferon-α/β signaling pathway GMV activity in the out-of-sample stimulated CD4 T cells matched the inferred activity in the in-sample CD4 T cells (Fig. 4a). To perform a more systematic comparison of the inferred latent space between out-of-sample and in-sample cells, we used the differential BF procedure (Methods) between (1) stimulated in-sample cells and control cells for a given cell type (model trained with the whole dataset) and (2) stimulated out-of-sample cells and control cells for the same cell type (model trained with one cell type/condition pair left out), and checked the amount of overlaps in the top 50 differentially activated GMVs (Fig. 4b). The results suggested consistency between the in-sample and out-of-sample differentially activated GMVs, with an average 72% overlap. To further evaluate the capacity of data reconstruction, we measured the $R^2$ between the original and decoded data in the in-sample and out-of-sample settings (Fig. 4c). We found that the $R^2$ decreases only marginally in the out-of-sample setting, confirming the ability of the model to generalize to unseen data produced in a similar experimental setting.

For the technical generalization test, we again checked that the interferon-α/β signaling pathway GMV activity distribution of study B encoded control CD4 T cells matched that of study A.
control CD4 T cells (Fig. 4d). We also investigated whether the top 50 differential GMVs of each cell type in a "one-vs-rest" differential setting for the control cells of study A overlapped with a similar procedure performed on the control cells of study B (Fig. 4e). We found that on average 67% of the top 50 differentially activated GMVs for study A overlap with those of study B, showing that the model can generalize across studies unseen at the time of training.

We then asked whether the model can use the inferred latent space to accurately reconstruct the original expression profiles of both studies. We found that the $R^2$ between original and reconstructed cells of study B, although lower than those for study A, improves upon the baseline correlation between the expression profiles of study A vs study B for most of the cell types (Fig. 4f).

**Discussion**

In this study, we introduced VEGA, a novel VAE architecture with a decoder inspired by known biology to infer the activity of various gene modules at the level of individual cells. By encoding single-cell transcriptomics data into an interpretable latent space specified a priori, our method provides a fast and efficient way of analyzing the activity of various biological abstractions in different contexts. In contrast, previous approaches used a posteriori interpretations of the latent variables to infer modules. VEGA’s flexibility in the specification of the latent space paves the way for analyzing the activity of biological modules such as pathways, transcriptional regulators, and cell type-specific modules. We illustrated how VEGA could be used to simultaneously investigate both cell type and cell state of cell subpopulations, in both control and experimentally perturbed conditions. Additionally, the weights of decoder connections provide direct interpretability of the relationship between the latent variables and the original features. For example, the decoder’s weights could be used to contrast interaction confidence in inferred GRNs or to rank genes by their importance in a certain biological module in a data-driven way. We further note that it was possible to modify VEGA’s architecture, following the same rationale as widely-used scVI9 and linear scVI13, such that it could handle count data in place of normalized expression profiles (SFig. 8).
The clear limitations of the current architecture resides in the sparse, single-layer decoder of the model. In fact, such an architectural design prevents the further improvement of generalizability and robustness. As a consequence, the generative capacity of VEGA is limited. For example, while VEGA theoretically could be used for interpretable response prediction using latent vector arithmetic in a similar fashion to scGen\textsuperscript{10}, VEGA’s limited generative capacity sacrifices predictive performance for biological interpretability of the latent space. We believe advanced insights in network biology, e.g., multi-layer GRNs that can describe regulatory machinery more comprehensively, could alleviate these limitations. This would open the possibility to perform targeted, in-silico activation, and repression of biological programs on specific cell populations to study its effect on development or disease progression. On the other hand, hard-coded connections of the linear decoder do not leave any room for correcting prior knowledge about gene modules when the context requires it, as is the case in other latent variable models such as f-scLVM\textsuperscript{12}. In fact, prior biological knowledge obtained from existing databases like MSigDB can be incomplete or not context-specific as additional unannotated genes can play an important role in certain gene modules. In parallel to our work on VEGA, Rybakov et al.\textsuperscript{33} introduced a regularization procedure to incorporate prior knowledge from gene annotation databases via a penalty term on the weights of the linear decoder. We demonstrated that VEGA performs comparatively to their interpretable autoencoder (SFig. 9), and that their approach is complementary to the unique attributes of VEGA and can be used to recover missing gene-GMV links in a data-driven fashion (SFig. 10).

In summary, we found VEGA useful for understanding the response of specific cell type populations to different perturbations, providing interpretable insights on biological module activity. The variational aspect of VEGA provides an advantage for addressing providing interpretable insights on biological module activity. The and can be used to recover missing gene-GMV links in a data-

Methods

The VEGA architecture. VEGA is a deep generative VAE that aims at maximizing the likelihood of a single-cell dataset X under a generative process\textsuperscript{3,10}, described as:

\[
p(X|\theta) = \int p(X|Z, \theta)p(Z|\theta)dZ.
\]  

(1)

with \( \theta \) being the learnable parameters of a neural network. VEGA uses a set of latent variables \( Z \) that explicitly represent sets of genes (gene modules), such as pathways, GRNs, or cell type marker sets. To enforce the VAE to interpret a dataset from the viewpoint of a set of gene modules, VEGA’s decoder part is made up of a single, masked, linear layer. Specifically, the connection of this layer, between latent node \( Z_j \) and gene features, are specified using a binary mask \( M \) in which \( M_{ij} \) is true if gene \( i \) is a member of a gene module \( j \) and false otherwise. We refer to each latent variable \( Z_j \) as a GMV since each provides a view of the data constrained to the subset of genes for a distinct gene module \( j \). During training, gradients associated with masked (false) weights are “zeroed out” such that backpropagation only applies to weights originating from a user-supplied given gene set. Additionally, the weights of the decoder are constrained to be nonnegative (\( w \geq 0 \)) to maintain interpretability as to the directionality of gene module activity.

Having explicitly specified the connections between genes and latent variables in the decoder of VEGA (generative part), we incentivize that the latent space represents a biological module activity interpretation of the data. We choose to model the GMVs as a multivariate normal distribution, parametrized by our inference network with learnable parameters \( \phi \). As such, the distribution of the \( Z \) latent variables can be expressed as:

\[
q(Z|X, \phi) = N(\mu(\theta(X)), \Sigma(\theta(X)));
\]  

\( q(Z|X, \phi) \) is the relative abundance of cells in group \( s \), and the integrals are approximated with direct Monte Carlo sampling. Similarly to Lopez et al.\textsuperscript{9}, assuming cells are independent, we can compute the average Bayes factor across many cell pairs randomly sampled from each group respectively. This helps us decide whether a GMV is activated at a higher frequency in one group or the other. Through the paper, we consider GMV’s to be significantly differentially activated if the absolute value of \( K \) is greater than 3 (equivalent to an odds ratio of \( \approx 20\))\textsuperscript{20},

\[
K = \log \frac{p(H_1|x_s, x_b)}{p(H_0|x_s, x_b)}.
\]  

\( H_1 \) and \( H_0 \) are two different hypotheses, e.g., different treatment conditions, or two mutually exclusive hypotheses:

\[
H_1^s := E_s[Z^s_x] > E_s[Z^b_x] \quad \text{vs.} \quad H_0^s := E_s[Z^s_x] \leq E_s[Z^b_x].
\]  

\( K \) can intuitively be seen as testing whether a cell has a higher mean GMV activation than another, the expectation representing empirical frequency. We evaluate the most probable hypothesis by studying the log-Bayes factor \( K \) defined as:

\[
K = \log \frac{p(H_1|x_s, x_b)}{p(H_0|x_s, x_b)}.
\]  

Here, the sign of \( K \) tells us which hypothesis is more likely, and the magnitude of \( K \) encodes a significance level. Having access to the conditional posterior distribution \( p(Z|X) \) over the GMV’s activation (the encoding part of VEGA), we can approximate each hypothesis’ probability distribution as:

\[
p(H_j^s|x_s, x_b) \approx \sum s p(s) \int_{\text{sup}(j, s)} \int_{\text{sub}(j, s)} p(z_j > z_j^b)dz_j dz_j^b.
\]  

where \( p(s) \) is the relative abundance of cells in group \( s \), and the integrals are approximated with direct Monte Carlo sampling.
Datasets and preprocessing

Kang et al. dataset. The Kang et al.17 dataset consisted of two groups of PBMCs, one control and one stimulated with interferon-β. We chose to use the same preprocessing steps as described by scGen authors16, using the Scanpy package35. Briefly, cells were annotated using the maximum correlation to one of the eight original cell type clusters identified, using an average of the top 20 cluster genes. Megakaryocytes were removed due to uncertainty about their annotation. Then data were filtered to remove cells with less than 500 genes expressed and genes expressed in five or less cells, using the scanpy.pp.filter_genes() and scanpy.pp.filter_cells() functions. Count per cells were then normalized and log-transformed using the scanpy.pp.normalize_per_cell() and scanpy.pp.log1p() functions, and we selected the top 6998 highly variable genes with scanpy.pp.highly_variable_genes(), resulting in a final dataset of 18,868 cells. Raw data is available at GSE96583. We used the same preprocessing functions for the rest of the datasets unless specified otherwise.

Zheng et al. dataset. The Zheng et al.32 dataset consists of 3K PBMCs from a healthy donor. After filtering the cells, the count per cells were normalized and log-transformed. We then subset the genes to use the same 6998 genes of the Kang et al. PBMC dataset. The final dataset has 2623 cells and 6998 genes. Raw data are available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/bmmc3k.

MIX-seq dataset. The MIX-seq19 datasets were obtained from https://figshare.com/s/139f64b495dea9d88c70, and we used the data from experiment 3 to have enough cells to carry a smooth training of our model. For the gene regulatory network analysis of GBM cells, we derived an ARACNe16,38 network from bulk RNA-Seq used in the MIX-Seq analysis part of this study. For the gene regulatory network, we ranked genes according to their FDR and used GSEA39 using the ge hypey package function gseapy.prerank() with the following settings: a minimum gene set size min_size=5, a maximum gene set max_size=1000, and a number of permutations permutation_num=1000. We ranked gene sets according to their FDR and considered significant hits when FDR ≤ 0.05. When the FDR returned by GSEA was equal to 0, we replaced it with 1e-5 (to avoid math error when taking the logarithm).

Batch correction comparison. To assess batch information integration in VEGA’s latent space, we compared the average silhouette scores on batch labels from the Shekhar et al. retina dataset (of 1) PCA with 50 principal components (computed using scanpy.pp.pca() function), (2) linear scVI13 as implemented in the scanpy.pp.linear() function, (3) VEGAs with following parameters: AnnData object setup with batch_key=batch, model initialized with n_hidden=800, n_layers=2, dropout_rate=0.2, n_latent=677, training performed with max_epochs=300, early_stopping=True, lr=5e-4, train_size=0.8, early_stopping_patience=300. VEGAs with following parameters: VEGAs object setup with batch_key=batch, model initialized using the REACTOME pathway database with three extra FC nodes to initialize the latent space and the same training hyperparameters as linear scVI.

Evaluation metrics. Silhouette scores were calculated to evaluate the separation of cell types and states in the latent space of our model. We used Euclidean distance in the latent space to compute the silhouette coefficient of each cell i defined as:

$$\text{silhouette}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where a(i) and b(i) are respectively the mean intra-cluster distance and the mean nearest-cluster distance for cell i. We used either the stimulation or cell type labels from Kang et al.17 to assess the biological relevance of the latent space of our model. The sklearn package42 silhouette_score() implementation was used for computation. For computing correlations throughout the paper, we used the function numpy.corrcoef() from the Numpy package43.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All of the datasets analyzed in this manuscript are publicly available. Please see the section Datasets and preprocessing for Methods for details. These datasets are also downloadable at https://github.com/LucasESBS/vega-reproducibility.

Code availability

The package and API for VEGA is available at https://github.com/LucasESBS/vega/tree/vega_dev44. The code and data to reproduce the results of this manuscript is available at https://github.com/LucasESBS/vega-reproducibility.

Received: 11 January 2021; Accepted: 13 September 2021; Published online: 28 September 2021

References

1. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620 (2015).
2. Hinton, G. E. Reducing the dimensionality of data with neural networks. Science 313, 504–507 (2006).
3. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. Proc. Mach. Learn. Res. 27, 37–49 (2012).
4. Wang, D. & Gu, J. VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Genomics, Proteomics Bioinformatics 16, 320–331 (2018).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-26017-0 | www.nature.com/naturecommunications
5. Geddes, T. A. et al. Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. *BMC Bioinformatics* **20**, 60 (2019).
6. Ersal Duek, C., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
7. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at arXiv:1312.6114 [cs, stat] (2014).
8. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* **23**, 80–91 (2018).
9. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
10. Lotfollahi, M., Wolf, A. F. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
11. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
12. Buettner, F., Pratapayani, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. fscLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
13. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
14. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
15. James, G., et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
16. Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**, 8 (2006).
17. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* **36**(January), 89–94 (2018). Number: 1 Publisher: Nature Publishing Group.
18. Mellor, A. L., Lemos, H. & Huang, L. Indoleamine 2,3-dioxygenase and tolerance: where are we now? *Front. Immunol.* **8**, 1360 (2017).
19. Sorgdrager, F. J. H., Naude, P. J. W., Kema, I. P., Nolen, A. E. & De Deyn, P. P. Tryptophan metabolism in inflammmation: from biomarker to therapeutic target. *Front. Immunol.* **10**, 2565 (2019).
20. Kass, R. E. & Raferty, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
21. McFarland, J. M. et al. Multiplexed single-cell transcriptomic response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
22. Liberron, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2017).
23. Kurata, K. et al. Growth arrest by activated BRAF and MEK inhibition in human anaplastic thyroid cancer cells. *Int. J. Oncol.* **49**, 2303–2308 (2016).
24. Joshi, M., Rice, S. J., Liu, X., Miller, B. & Belani, C. P. Trametinib with or without vemurafenib in BRAF mutated non-small cell lung cancer. *PloS ONE* **10**, e0118210 (2015).
25. Lulli, D., Carbone, M. L. & Pastore, S. The MEK inhibitors trametinib and cobimetinib induce a type I interferon response in human keratinocytes. *Int. J. Mol. Sci.* **18**, 2227 (2017).
26. Spitz, F. & Furlong, E. M. Transmission factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
27. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
28. Carro, M. S. et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
29. Darmanis, S. et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Reports* **21**, 1399–1410 (2017).
30. Lu, R. Q. et al. Common developmental requirement for olig function indicates a motor neuron/oligodenodcyte connection. *Cell* **109**, 75–86 (2002).
31. Field, A. R. et al. Structurally conserved primate LncRNAs are transiently expressed during human cortical differentiation and influence cell-type specific genes. *Stem Cell Reports* **12**, 245–257 (2019).
32. Zheng, G. X., et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
33. Rybakov, S., Lotfollahi, M., Theis, F. J. & Wolf, A. F. Learning interpretable latent autoencoder representations with annotations of feature sets. Preprint at bioRxiv https://doi.org/10.1101/2020.02.02.041182 (2020).
34. Yeung, S., Kannan, A., Daphins, Y. & Fei-Fei, L. Tackling over-pruning in variational autoencoders. Preprint at arXiv: 1706.03643 (2017).
35. Wolf, A. F., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
36. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e130 (2016).
37. Cao, K. L. et al. Ensembler 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
38. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNE-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).
39. Ding, H. et al. Quantitative assessment of protein activity in ovarian tissues and single cells using the metaViper algorithm. *Nat. Commun.* **9**, 1471 (2018).
40. McNinnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv: 1802.03426 (2020).
41. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
43. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
44. Seneing, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. vegi. https://doi.org/10.5281/zenodo.5338892 (2021).

**Acknowledgements**

L.S. was supported by the Schmidt Futures Foundation SF 857 and by the National Institute Of Mental Health of the National Institutes of Health award R01MH120295. J.M.S. was supported by a grant 5R01GM106031 from the NIGMS. J.S. and H.D. were supported by a grant from the Chan-Zuckerberg Initiative’s Human Cell Atlas portals project. H.D. was supported by a gift from Seagate Technology. J.S. was supported by grant GCIR-06673-C from the California Institute for Regenerative Medicine’s Center of Excellence for Stem Cell Genomics. The authors would like to thank Dr. David Haussler and Dr. Sofi Salama for their support. L.S. would also like to thank David Parks for the useful feedback during the early development of the method. We also would like to thank Dr. Maximilian Haasler for the feedback on the manuscript.

**Author contributions**

L.S. and I.A. conceived the idea. L.S. implemented the method and gathered the data. L.S. and I.A. performed the analysis. H.D. and J.S. supervised the research. All authors contributed to the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26017-0.

**Correspondence** and requests for materials should be addressed to Hongxu Ding or Joshua Stuart.

**Peer review information** *Nature Communications* thanks Gokcen Eralan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.