Dhaka: Variational Autoencoder for Unmasking Tumor Heterogeneity from Single Cell Genomic Data

Sabrina Rashid,[1] Sohrab Shah[2,3,4], Ziv Bar-Joseph[5], Ravi Pandya[6*]

[1] Computational Biology Department, Carnegie Mellon University, Pittsburgh, USA
[2] Department of Computer Science, University of British Columbia, Vancouver, Canada
[3] Department of Molecular Oncology, BC Cancer Agency, Vancouver, Canada
[4] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
[5] Machine Learning Department and Computational Biology Department, Carnegie Mellon University, Pittsburgh, USA
[6] Microsoft Research, Redmond, USA
*To whom correspondence should be addressed; E-mail: ravip@microsoft.com

Abstract. Intra-tumor heterogeneity is one of the key confounding factors in deciphering tumor evolution. Malignant cells exhibit variations in their gene expression, copy numbers, and mutation even when originating from a single progenitor cell. Single cell sequencing of tumor cells has recently emerged as a viable option for unmasking the underlying tumor heterogeneity. However extracting features from single cell genomic data in order to infer their evolutionary trajectory remains computationally challenging due to the extremely noisy and sparse nature of the data. Here we describe 'Dhaka', a variational autoencoder method which transforms single cell genomic data to a reduced dimension feature space that is more efficient in differentiating between (hidden) tumor subpopulations. Our method is general and can be applied to several different types of genomic data including copy number variation from scDNA-Seq and gene expression from scRNA-Seq experiments. We tested the method on synthetic and 6 single cell cancer datasets where the number of cells range from 250 to 6000 for each sample. Analysis of the resulting feature space revealed subpopulations of cells and their marker genes. The features are also able to infer the lineage and/or differentiation trajectory between cells greatly improving upon prior methods suggested for feature extraction and dimensionality reduction of such data.

Keywords: Single cell genomic data, Neural Networks, Tumor Heterogeneity, Differentiation and/or lineage trajectory

Supporting info and Software: `https://github.com/MicrosoftGenomics/Dhaka`

Introduction

Tumor cells are often very heterogeneous. Typical cancer progression consists of a prolonged clinically latent period during which several new mutations arise leading to changes in gene expression and DNA copy number for several genes [1], [2], [3]. As a results of such genomic variability, we often see multiple subpopulations of cells within a single tumor.

The goal of effective cancer treatment is to treat all malignant cells without harming the originating host tissue. Clinical approaches should thus take into account the underlying evolutionary structure in order to identify treatments that can specifically target malignant cells while not affecting their normal cell of origin. It is also important to determine if the ancestral tumor clones eventually disappear (chain like evolution) or if several genotypically different clones of cells evolved in parallel (branched evolution) [1]. Tumors resulting from these two evolutionary trajectories respond differently, and ignoring the evolutionary process when determining treatment can lead to therapy resistance and possible cancer recurrence. Thus, characterization of the hidden subpopulations and their underlying evolutionary structure is an important issue for both the biological understanding and clinical treatment of cancer. Prior studies have mainly relied on bulk sequencing to investigate tumor evolution [4], [5]. In such experiments thousands of cells are sequenced together, which averages out the genomic characterisitics of the individual cells making it hard to infer these sub-populations. More recently, single cell sequencing has emerged as a useful tool to study such cellular heterogeneity [6], [7], [8], [9].

While single cell data is clearly much more appropriate for addressing tumor heterogeneity and evolution, it also raises new computational and experimental challenges. Due to technical challenges (for example, the low quantity of genetic material and the coverage for each of the cells sequenced) the resulting data is often very noisy and sparse with many dropout events [10],[11]. These issues affect both, scRNA-Seq and scDNA-Seq experiments which are used for copy number and mutation estimation. Given these issues, it remains challenging to identify meaningful features that can accurately characterize the single cells in terms of their clonal identity and differentiation state. To address this, several methods have been proposed to transform the observed gene expression or copy number profiles in order to generate features that are more robust for down stream analysis. However, as we show below, many of the feature transformation techniques that are usually applied to genomic data fail to identify the sub populations and their trajectories. For example, while t-SNE [12] and diffusion maps [13] are very successful in segregating cells between different tumor samples, they are less successful when trying to characterize the evolutional trajectories of a single tumor. Single cell clustering algorithms, including SNN-cliq [14] which constructs a shared k-nearest neighbor graph across all cells and then finds maximal cliques and PAGODA [15] which relies on prior set of annotated genes to find transcriptomal heterogeneity, can successfully distinguish between different groups of cells in a dataset. However, such methods are not designed for determining the relationship between the detected clusters which is the focus of tumor evolutionary analysis. In addition, most current single cell clustering methods are focused on only one type of genomic data (for example scRNA-Seq) and do not work well for multiple types of such data.

Another direction that has been investigated for reducing the dimensionallity of scRNA-Seq data is the use of neural networks (NN) [16], [17]. In Lin *et al.* [16], the authors used prior biological knowledge including protein-protein and protein-DNA interaction to learn the architecture of a NN and to subsequently project the data to a lower dimensional feature space. Unlike these prior approaches, which were supervised, we are using neural networks in a completely unsupervised manner and so do not require labeled data as prior methods have. Specifically, in our software 'Dhaka' we have developed a variational autoencoder that combines Bayesian inference with unsupervised deep learning, to learn a probabilistic encoding of the input data. Our autoencoder method can be

used for analyzing different types of genomic data. Specifically, in this paper we have analyzed 4 scRNA-Seq and 2 scDNA-Seq datasets. We used the variational autoencoder to project the expression and copy number profiles of tumor populations and were able to reconstruct their trajectories even for noisy sparse datasets with very low coverage. These results highlight the effectiveness of the our method in extracting important biological and clinical information from cancer samples.

Methods

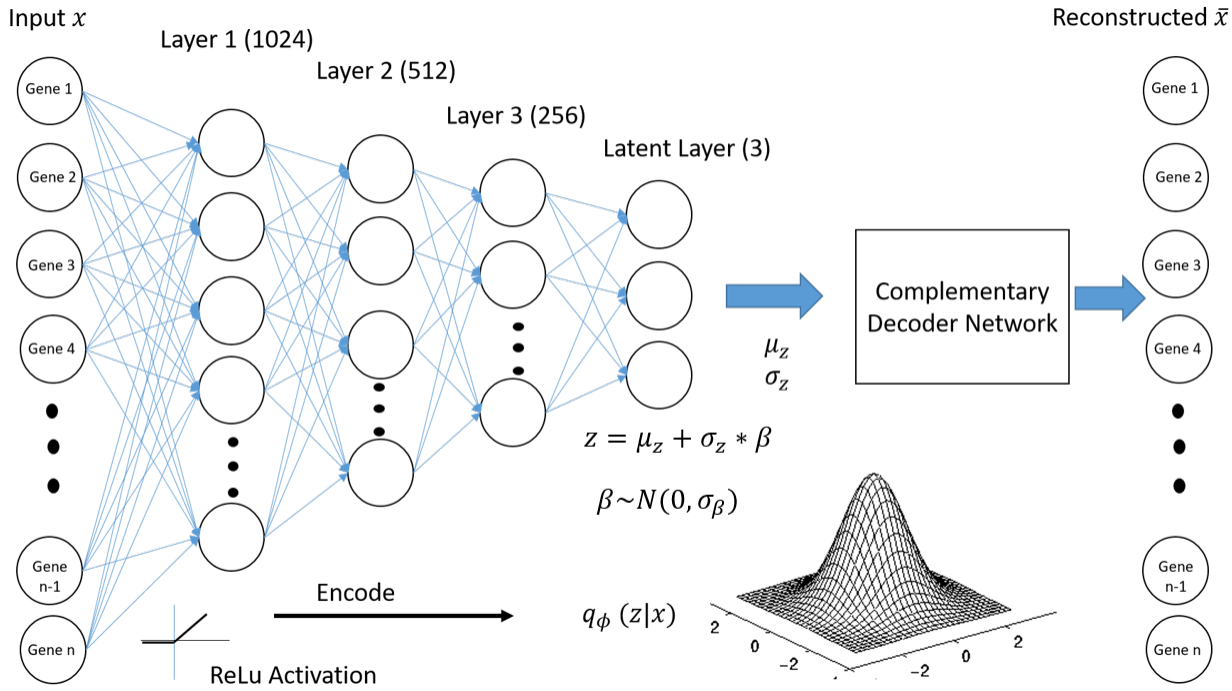

Fig. 1. Structure of the variational autoencoder used in Dhaka.

Variational autoencoder

We used a variational autoencoder to analyze single cell genomic data. For this, we adapted an autoencoder initially proposed by [18]. *Autoencoders* are multilayered perceptron neural networks that sequentially deconstruct data (x) into latent representation (z) and then use these representations to reconstruct outputs that are similar (in some metric space) to the inputs. The main advantage of this approach is that the model learns the best features and input combinations in a completely unsupervised manner. In *variational autoencoders (VAE)* unsupervised deep learning is combined with Bayesian inference. Instead of learning an unconstrained representation of the data we impose a regularization constraint. We assume that the latent representation is coming from a probability distribution, in this case a spherical Gaussian ($N(\mu_z, \sigma_z)$). The intuition behind such representation for single cell data is that the heterogeneous cells are actually the result of some underlying biological process leading to the observed expression and copy number data. These processes are modeled here as distribution over latent space, each having their distinct means and variances. Hence the autoencoder actually encodes not only the mean (μ_z) but also the variance(σ_z) of this Gaussian distribution. The latent representation (z) is then sampled from the learned posterior

distribution $q_\phi(z|x) \sim N(\mu_z, \sigma_z I)$. Here ϕ are the parameters of the encoder network (such as biases and weights). The sampled latent representation is then passed through a similar decoder network to reconstruct the input $\bar{x} \sim p_\theta(x|z)$, where θ are the parameters of the decoder network. Although the model is trained holistically, we are actually interested in the latent representation z of the data since it represents the key information needed to accurately reconstruct the inputs.

Model structure: Fig. 1 presents the structure of the autoencoder used in this paper. The input layer consists of nodes equal to the number of genes we are analyzing for each cell. We are introducing nonlinearity in the model by using Rectified Linear unit(ReLu) activation function. We have used three intermediate layers with 1024, 512, and 256 nodes and a 3-node latent layer.

The size of the latent dimension (i.e. the representation we extract from the model) is a parameter of the model. As we show in Results, for the data analyzed in this paper three latent variables are enough to obtain accurate separation of cell states for both the expression and copy number datasets. Increasing this number did not improve the results and so all figures and subsequent analysis are based on this number. However, the method is general and if needed can use more or less nodes in the latent layer.

All datasets we analyzed had more than 5k genes and the reported structure with atleast 1024 nodes in the first intermediate layer (Fig. 1) was sufficient for them. We used three intermediate layers to gradually compress the encoding to a 3 dimensional feature space. We have also compared two different structures of autoencoders with the proposed three intermediate layers vs one intermediate layer in the Results section. More intermediate layers leads to better performance but slightly increases the runtime.

Learning: To learn the parameters of the autoencoder, ϕ and θ, we need to maximize $log(p(x|\phi, \theta))$, the log likelihood of the data points x, given the model parameters. The marginal likelihood $log(p(x))$ is the sum of a variational lower bound [18] and the Kulback-Leibler (KL) [19] divergence between the approximate and true posteriors.

$$log(p(x)) = L(\phi, \theta; x) + D_{KL}(q_\phi(z|x)||p_\theta(z|x))$$

The likelihood L can be decomposed as following:

$$L(\phi, \theta; x) = E_{z \sim q_\phi(z|x)}[log(p_\theta(x|z))] - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

The first term can be viewed as the typical reconstruction loss intrinsic to all autoencoders, the second term can be viewed as the penalty for forcing the encoded representation to follow the Gaussian prior (the regularization part). We then use 'RMSprop', which relies on a variant of stochastic minibatch gradient descent, to minimize $-L$. In 'RMSprop', the learning rate weight is divided by the running average of the magnitudes of recent gradients for that weight leading to better convergence [20]. Detailed derivation of the loss computation can be found in [18].

An issue in learning VAE with standard gradient descent is that gradient descent requires the model to be differentiable, which is inherently deterministic. However, in VAE the the fact that we sample from the latent layer makes the model stochastic. To enable the use of gradient descent in our model, we introduce a new random variable β. Instead of sampling z directly from the $N(\mu_z, \sigma_z I)$, we set

$$z = \mu_z + \sigma_z * \beta$$

Where β is the Gaussian noise, $\beta \sim N(0, \sigma_\beta)$. Using β we do not need to sample from the latent layer and so the model is differentiable and gradient descent can be used to learn model parameters [21]. σ_β is the standard deviation of the Gaussian noise and is an input parameter of the model.

Results

Quantitative validation with simulated dataset

We first performed simulation analysis to compare the Dhaka method with prior dimsionallity reduction methods that have been extensively used for scRNA-Seq data: t-SNE [12] and PCA [22]. We generated 2 simulated datasets with 3K genes and 500 cells. To generate the simulated data we followed the method described in [23]. In each dataset, cells are generated from five different clusters with 100 cells each. Both datasets contain a mix of noisy genes (that are not related to the cluster the cell belongs to) and genes with cluster specific expression. In the first dataset, 500 of the 3K genes are noisy while in the 2nd 2500 of the 3K are noisy. We have used a Gaussian Mixture Model to cluster the reduced dimension data and Bayesian Information criterion (BIC) to select the number of clusters. We next compute the Adjusted Random Index (ARI) metric to determine the quality of resulting clustering for each dimensionality reduction method. Fig 2 (first column) shows the result of our autoencoder, PCA, and t-SNE projection for the 500 noisy genes simulated data. As can be seen, only the Dhaka autoencoder correctly identifies all 5 clusters (Fig 2g), whereas PCA identifies only 1 and t-SNE only 3. The ARI for the autoencoder projection is 0.98 indicating almost perfect match to the original assignments whereas it is 0 for PCA and 0.061 for t-SNE.

When the number of noisy genes increases to 2500 the Dhaka autoencoder still identifies 4 clusters, whereas t-SNE only identifies 2. PCA also identifies 4 clusters but with a much lower ARI score of .16 (Fig 2 second column). Although the autoenoder drops one cluster it can still differentiate others very well resulting in a high ARI score of 0.71. This corroborates the robustness of the method used in Dhaka.

We have also compared two different structures of the autoencoder (structure 1: $Input \rightarrow 1024 \quad nodes \rightarrow 512 \quad nodes \rightarrow 256 \quad nodes \rightarrow 3 \quad latent \quad dims$, structure 2: $Input \rightarrow 1024 \quad nodes \rightarrow 3 \quad latent \quad dims$ in terms of ARI and runtime (Table 1) on the sythetic data with 2500 noisy genes. The VAE structure 1 (Fig. 1) gives the best ARI score and both the VAE structures are faster than t-SNE.

	Structure 1	Structure 2	t-SNE
ARI	0.71	0.5	0.27
Runtime (s)	3.43	2.13	3.57

Table 1. Comparison between structures of autoencoders and t-SNE

Gene expression data

We have next tested the method on four single cell RNA-seq tumor datasets: i) Oligodendroglioma [7], ii) Glioblastoma [24], iii) Melanoma [25], and iv) Astrocytoma [6]. We discuss the first three below and the fourth in the appendix.

Analysis of Oligodendroglioma data

Oligodendrogliomas are a type of glioma that originate from the oligodendrocytes of the brain or from a glial precursor cell. In the Oligodendroglioma dataset the authors profiled six untreated Oligodendroglioma tumors resulting in 4347 cells and $23K$ genes. The dataset is comprised of both malignant and non-malignant cells. Copy number variations (CNV) were estimated from the $log2$ transformed transcript per million RNA-seq expression data . The authors then computed two metrics, lineage score and differentiation score by comparing pre-selected 265 signature genes' CNV profile for each cell with that of a control gene set. Based on these metrics, the authors determined that the malignant cells are composed from two subpopulations, oligo-like and astro-like, and that both share a common lineage. The analysis also determined the differentiation state of each cell.

Here we are using the RNA-seq expression data directly skipping the CNV analysis. With only three latent dimensions our algorithm successfully separated malignant cells from non malignant

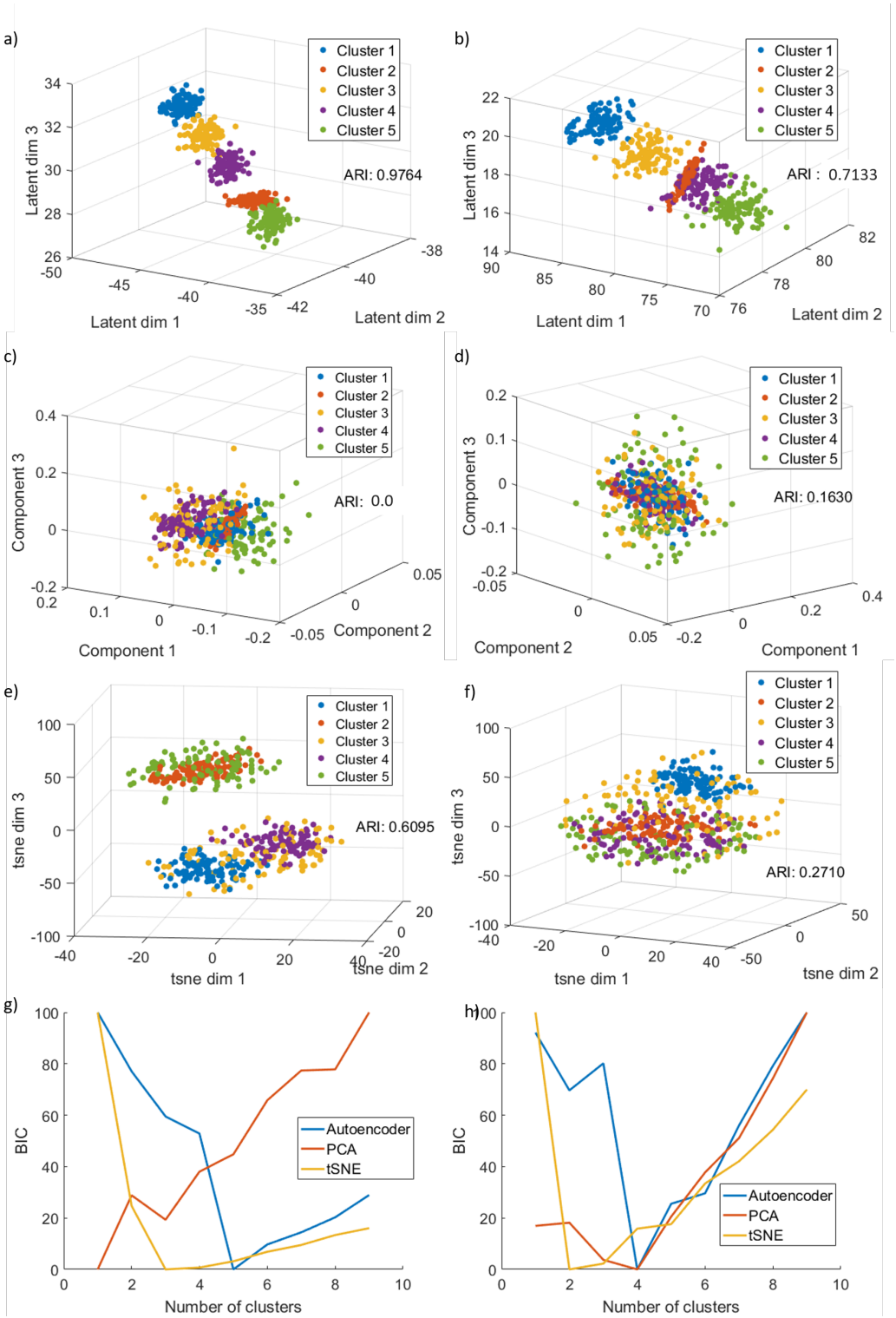

Fig. 2. Comparison of the Dhaka method with t-SNE and PCA on two simulated datasets. First column: result on simulated dataset with 500 noisy genes (17% of total genes). Second column: result on simulated dataset with 2500 noisy genes (83% of total genes). a),b) Autoencoder output. c),d) t-SNE output. e),f) PCA output. g),h) Plot of BIC calculated from fitting Gaussian Mixture Model to the data to estimate number of clusters. The number with lowest BIC is considered as the estimated number of clusters in the data.

microglia/macrophage cells (Fig. 3a). We next analyzed the malignant cells only using their relative expression profile (see appendix), to identify the different subpopulations and the relationship between them. Fig. 3b-c show the projected autoencoder output, where we see two distinct subpopulations originating from a common lineage, thus recapitulating the finding of the original paper. The autoencoder was not only able to separate the two subpopulations, but also to uncover their shared glial lineage. To compare the results with the original paper, we have plotted the scatter plot with color corresponding to lineage score (Fig. 3b)and differentiation score (Fig. 3c) from [7]. We can see from the figure that the autoencoder can separate oligo-like and astro-like cells very well by placing them in opposite arms of the v-structure. In addition, figure Fig 3c shows that most of the cells with stem like property are placed near the bifurcation point of the v-structure.

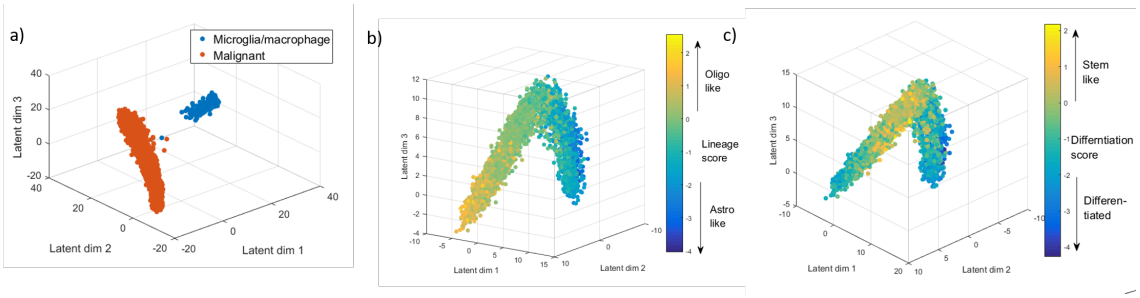

Fig. 3. Oligodendroglioma dataset.a) Autoenocder projection separating malignant cells from non malignant microglia/macrophage cells. b)-c) Autoencoder output from relative expression profile of malignant cells using 265 signatures genes. b) Each cell is colored by their assigned lineage score which differentiates the oligo-like and astro-like subpopulations. c) Each cell is colored by their assigned differentiation score, which shows that most stem like cells are indeed placed near the bifurcation point.

The analysis discussed above was based on the 265 signature genes that were reported in the original paper. We next tested whether a similar structure can be learned from auto selected genes, instead of using these signature genes. Fig. 4a shows the autoencoder projection using 5000 auto-selected genes based on $\bar{A}$ score (see appendix). As we can see from Fig. 4a, the autoencoder can learn similar structure without the need for supervised prior knowledge. We also compared the autoencoder output for this data to t-SNE and PCA (Fig. 4b-c). As can be seen, PCA can separate the oligo-like and astro-like structure to some extent, but their separation is not as distinct as the autoencoder output. t-SNE can recover clusters of cells from the same tumor but completely fails to identify the underlying lineage and differentiation structure of the data.

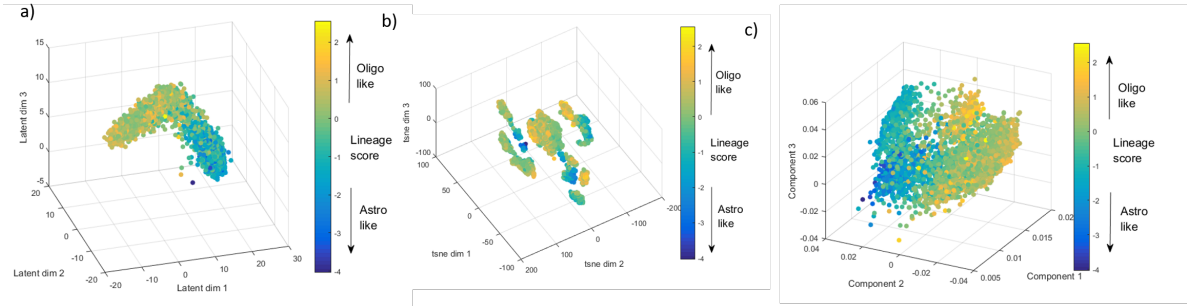

Fig. 4. Comparison of variational autoencoder with t-SNE and PCA. a) Autoencoder output using 5000 auto-selected genes colored by lineage score. b) t-SNE projection colored by lineage score. c) PCA projection colored by lineage score.

Robustness analysis: A key issue with the analysis of scRNA-Seq data is dropout. In scRNA-Seq data we often see transcripts that are not detected even though the particular gene is expressed, which is known as the 'dropout'. This happens mostly because of the low genomic quantity used for scRNA-Seq. We have tested the robustness of the autoencoder to dropouts in the Oligodendroglioma dataset. We tested several different dropout percentages ranging from 0 to 50% (Supporting Fig. 10a). Supporting Fig. 10c,e,g shows the histogram after artificially forcing 20%, 30%, and 50% more genes to be dropped out. Note that we can not go beyond 50% in our analysis since several genes are already zero in the original data. Supporting Fig. 10b,d,f,h shows the projection of the autoencoder after adding 0%, 20%, 30%, and 50% more drop out genes, respectively. We observe that when the additional drop out rate is 30% or less, the autoencoder can still retain the v-structure even though the cells are a bit more dispersed. At 50% we lose the v-structure, but the method can still separate oligo-like and astro-like cells even with this highly sparse data. We have also performed quantitative analysis of the drop out effect on the synthetic dataset with 2500 noisy genes. The Dhaka autoencoder is consistently more robust than both t-SNE and PCA. The details of the comparison can be found in Appendix (see Supporting Fig. 11).

Analysis of marker genes in the Oligodendroglioma dataset: We further investigated the autoencoder learned structure to discover genes that are correlated with the lineages. To obtain trajectories for genes in the two lineages of the Oligodendroglioma dataset, we first segmented the autoencoder projected output into 9 clusters using Gaussian mixture model (Fig 5). Clusters 1 to 4 correspond to the oligo branch and clusters -4 to -1 correspond to the astro branch, while cluster 0 represents the bifurcation point. After computing the average expression profile of genes in the oligo and astro branches, we performed two tailed t-test to identify differentially expressed (DE) genes among the group of cells in the two lineages. With adjusted p-value < 0.05, we find 1197 DE genes. We have also separately identified genes that are up regulated and down regulated in the two lineages (see list of genes here). Expression profiles of a few of these genes are shown in Fig. 5b)-e). Many of the genes we identified were known to be related to other types of cancers or neurological disorders, but so far have not been associated with Oligodendroglioma. For example, TFG which is upregulated in the oligo-branch was previously affiliated in neuropathy [26]. DDX39B gene is not directly related to cancer, but is found to be localized near genes encoding for tumor necrosis factor α and β [27]. Both HEXB and RGMA genes are up regulated in the astro-branch. These genes were previously identified in neurological disorders such as Sandhoff disease [28] and multiple sclerosis [29], respectively. Our analysis suggests that they are key players in the Oligodendroglioma pathway as well.

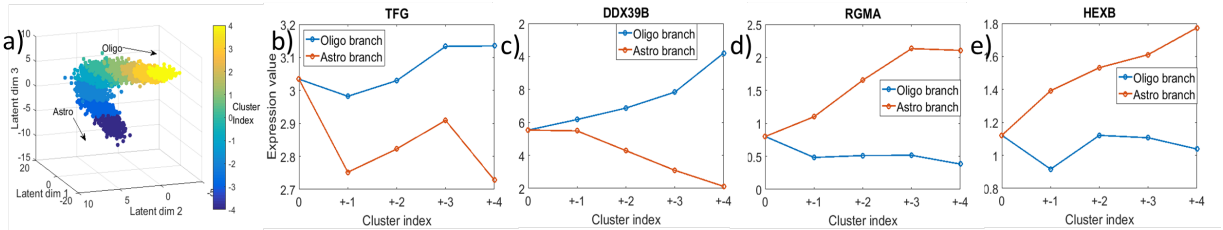

Fig. 5. New gene markers for astro-like and oligo-like lineages. a) Segmenting autoencoder projected output to 9 clusters. Clusters -4, -3,-2,-1 belongs to astro branch and clusters 1,2,3,4 belong to oligo branch. Cluster 0 represent the origin of bifurcation. b)-e) Expression profiles of couple of the top differntially expressed genes in the two lineages. b)-c) upregulated in the oligo-branch, d)-e) upregulated in the astro-branch.

Analysis of Glioblastoma data

The next dataset we looked at is the Glioblastoma dataset [24]. This dataset contains 420 malignant cells with $\sim$ 6000 expressed genes from six tumors. In this relatively small cohort of cells the authors did not find multiple subpopulations. However they identified a stemness gradient across the cells from all six tumors [24], meaning the cells gradually evolve from a stemlike state to a more differentiated state. When we applied the Dhaka autoencoder to the expression profiles of the malignant cells, the cells were arranged in a chain like structure 6a). To correlate the result with the underlying biology, we computed stemness score from the signature genes reported in the original paper (78 genes in total) [24]. The score is computed as the ratio of average expression of the stemness signature genes to the average expression of all remaining genes [24]. When we colored the scattered plot according to the corresponding stemness score of each cell, we see a chain like evolutionary structure where cells are gradually progressing form a stemlike state to a more differentiated state. As before, t-SNE and PCA projections (Fig. 6b-c), fail to capture the underlying structure of this differentiation process.

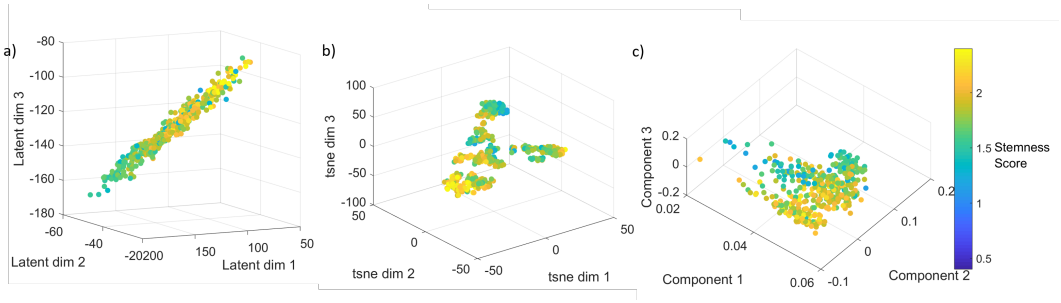

Fig. 6. Comparison of variational autoencoder with t-SNE and PCA on Glioblastoma dadaset. a) Autoencoder output using 5000 auto-selected genes colored by stemness score. b) t-SNE projection colored by stemness score. c) PCA projection colored by stemness score.

After learning the structures we also wanted to see whether we can identify new marker genes for the stemness to differentiated program. For this, we reduced the latent dimension to 1 (since we see almost linear projection). Next we computed Spearman rank correlation [30] of the 1D projection with every gene in the dataset. We have plotted a few of the top ranked positive (up regulated in the stemlike cells) and negative correlated genes (down regulated in the stemlike cells) (Fig. 7a). Despite the noisy expression profile, we do see a clear trend when a line is fitted (red). Among the discovered markers, TPT1 was identified as one of the key tumor proteins [31]. Both RPS27 and TPT1 were found to be significant in other forms of cancer, such as Melanoma [32] and prostate cancer [31] and our results indicate that they may be involved in Glioblastoma as well. Among the downregulated genes, CLU was identified in the original paper [24] to be affiliated in Glioblastoma pathway whereas CANX was previously not identified as a marker for Glioblastoma. A complete list of correlated marker genes can be found here.

Analysis of Melanoma data

The Melanoma cancer dataset [25] profiled 1252 malignant cells with $\sim 23k$ genes from 19 samples. The expression values are log2 transformed transcript per million. When we used the relative expression values of 5000 auto-selected genes (based on $\bar{A}$ score) to the Dhaka autoencoder we saw two very distinct clusters of cells, revealing the intra-tumor heterogeneity of the Melanoma samples (Fig. 8a). In the original paper, the authors identified two expression programs related to MITF and AXL genes that gives rise to a subset of cells that are less likely to respond to targeted therapy.

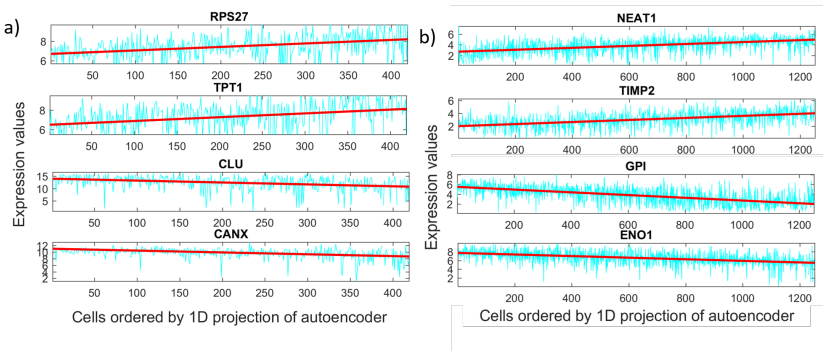

Fig. 7. New marker genes for: a) Glioblastoma differentiation program b) Melanoma MITF-AXL program.

The signature score for these program was calculated by identifying genesets correlated with these two programs. The authors identified a total of 200 signature genes. We computed MITF-AXL signature score by computing the ratio of average expression of the signature genes and average expression of all remaining genes. When we colored the scattered plot with the MITF-AXL score, we indeed see that the clusters correspond to the MITF-AXL program, with one cluster scoring high and the other scoring low for these signature genes. Again, such heterogeneity is not properly captured by t-SNE and PCA (Fig 8b-c).

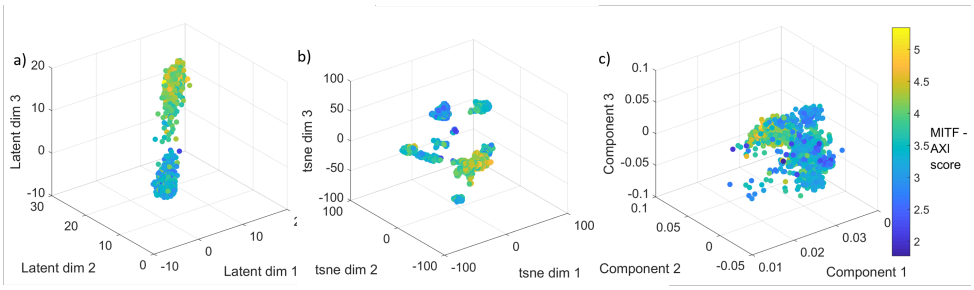

Fig. 8. Comparison of variational autoencoder with t-SNE and PCA on Melanoma dadaset. a) Autoencoder output using 5000 auto-selected genes colored by MITF-AXL score. b) t-SNE projection colored by MITF-AXL score. c) PCA projection colored by MITF-AXL score.

For this case too, we see almost a linear projection. To find new gene markers, we again computed 1D latent projection of the single cells and computed gene correlation. We have plotted a set of new marker genes both up and down regulated (Fig. 7b)). The NEAT1 is a non-coding RNA, which acts as a transcriptional regulator for numerous genes, including some genes involved in cancer progression [33]. TIMP2 gene plays a critical role in suppressing proliferation of endothelial cells and now we can see it is also relevant in the Melanoma cells [34]. Among the down regulated genes, GPI functions as tumor-secreted cytokine and an angiogenic factor, which is very relevant to any cancer progression [35]. The last correlated down regulated gene ENO1 is also known as tumor suppressor [36]. We have also looked whether the projection can recover some known gene marker dynamics or not. Four of the known gene markers are plotted in supporting Fig. 12 (in Appendix). A complete set of gene markers can be found here.

We have also analyzed another scRNA-seq tumor dataset, Astrocytoma. Due to space constraint we have moved the analysis to the Appendix.

Copy number variation data

To test the generality of the method we also tested Dhaka with copy number variation data. We used copy number profiles from two xenograft breast tumor samples (xenograft 3 and 4, representing two consecutive time points) [8]. 260 cells were profiled from Xenograft 3 and 254 from xenograft

4. Both of these datasets have around 20K genomic bin count. Cells were sequenced at a very low depth of 0.05X which results in noisy profiles. Copy numbers were estimated using a hidden Markov model [37]. When we analyzed the copy number profile for xenograft 3, the autoencoder identified 1 major cluster of cells and 1 minor cluster of cells (Fig 9a). The identified clusters agree with the phylogenetic reconstruction analysis in the original paper. Fig. 9b shows the copy number profiles of cells organized by phylogenetic analysis. Even though the copy number profiles are mostly similar in most parts of the genome, we do see that there is a small number of cells that have two copies (as opposed to one in the majority of cells, marked by red circle) in the x chromosome. The autoencoder was able to correctly differentiate the minor cluster of cells from the rest. Next we analyzed the xenograft 4 samples. The projected autoencoder output showed only 1 cluster which overlaps the major cluster identified for xenograft 3. We believe that the minor cluster from xenograft 3 probably did not progress further after serial passaging to the next mouse, whereas the major cluster persisted. This observation also agrees with the claim stated in the original paper [8] that after serial passaging only one cluster remained in xenograft 4 which is a descendant of the major cluster in xenograft 3.

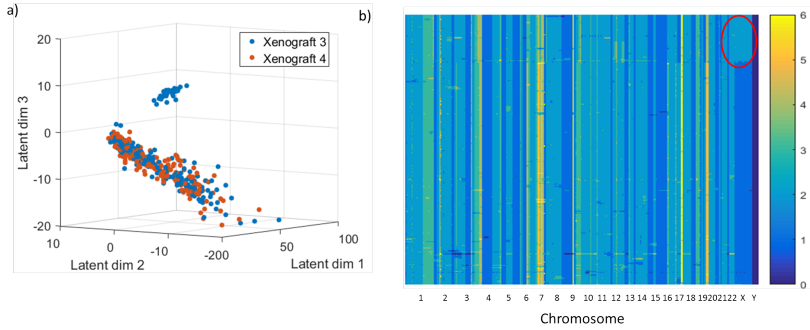

Fig. 9. Autoencoder ouptut of two xenograft breast tumor samples' copy number profile. a) Identification of two subpopulations of cells in xenograft 3 and one subpopulation in xenograft 4. b) Copy number profile of cells in xenograft 3 ordered by phylogenetic analysis, which shows that there are indeed two groups of cells present in the data.

Discussion

In this paper, we have proposed a new way of extracting useful features from single cell genmoic data. The method is completely unsupervised and requires minimal preprocessing of the data. Using our method we were able to reconstruct lineage and differentiation ordering for several single cell tumor samples. The autoencoder successfully separated oligo-like ans astro-like cells along with their differentiation status for Oligodendroglioma scRNA-Seq data and has also successfully captured the differentiation trajectory of Glioblastoma cells. Similar results were obtained for Melanoma and Astrocytoma. The method is general and can be applied to other type of genomic data as well. When applied to copy number variation data the method was able to identify heterogeneous tumor populations for breast cancer samples. The autoencoder projections have also revealed several new marker genes for the cancer types analyzed.

An advantage of the autoencoder method is its ability to handle dropouts. Several single cell algorithms require preprocessing to explicitly model the drop out rates. As we have shown, our method is robust and can handle very different rates eliminating the need to estimate this value.

While our focus here was primarily on the identification of sub populations and visualization, the latent representation generated by the autoencoder could be used in pseudotime ordering algorithms as well [38], [39]. These methods often rely on t-SNE/PCA as the first step and replacing these with the autoencoder is likely to yield more accurate results as we have shown. The variational autoencoder does not only clusters the cells, it can also represent an evolutionary trajectory, for example, the V structure for the Oligodendroglioma. Hence it can also be useful in phylogenetic analysis. Potential future work would focus on using this to identify key genes that align with the progression and mutations that help drive the different populations.

References

1. E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, *et al.*, "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution," *Science*, vol. 346, no. 6206, pp. 251–256, 2014.
2. N. Andor, T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley, "Pan-cancer analysis of the extent and consequences of intra-tumor heterogeneity," *Nature medicine*, vol. 22, no. 1, p. 105, 2016.
3. J.-W. Min, W. J. Kim, J. A. Han, Y.-J. Jung, K.-T. Kim, W.-Y. Park, H.-O. Lee, and S. S. Choi, "Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell rna-seq," *PloS one*, vol. 10, no. 8, p. e0135817, 2015.
4. N. E. Navin and J. Hicks, "Tracing the tumor lineage," *Molecular oncology*, vol. 4, no. 3, pp. 267–283, 2010.
5. H. G. Russnes, N. Navin, J. Hicks, and A.-L. Borresen-Dale, "Insight into the heterogeneity of breast cancer through next-generation sequencing," *The Journal of clinical investigation*, vol. 121, no. 10, p. 3810, 2011.
6. A. S. Venteicher, I. Tirosh, C. Hebert, K. Yizhak, C. Neftel, M. G. Filbin, V. Hovestadt, L. E. Escalante, M. L. Shaw, C. Rodman, *et al.*, "Decoupling genetics, lineages, and microenvironment in idh-mutant gliomas by single-cell rna-seq," *Science*, vol. 355, no. 6332, p. eaai8478, 2017.
7. I. Tirosh, A. S. Venteicher, C. Hebert, L. E. Escalante, A. P. Patel, K. Yizhak, J. M. Fisher, C. Rodman, C. Mount, M. G. Filbin, *et al.*, "Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma," *Nature*, vol. 539, no. 7628, pp. 309–313, 2016.
8. H. Zahn, A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio, and C. L. Hansen, "Scalable whole-genome single-cell library preparation without preamplification," *Nature methods*, vol. 14, no. 2, pp. 167–173, 2017.
9. A. Giustacchini, S. Thongjuea, N. Barkas, P. S. Woll, B. J. Povinelli, C. A. Booth, P. Sopp, R. Norfo, A. Rodriguez-Meira, N. Ashley, *et al.*, "Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia," *Nature medicine*, vol. 23, no. 6, pp. 692–702, 2017.
10. C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current state of the science," *Nature reviews. Genetics*, vol. 17, no. 3, p. 175, 2016.
11. C. Zong, S. Lu, A. R. Chapman, and X. S. Xie, "Genome-wide detection of single-nucleotide and copy-number variations of a single human cell," *Science*, vol. 338, no. 6114, pp. 1622–1626, 2012.
12. L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
13. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
14. C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.
15. J. Fan, N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J.-B. Fan, K. Zhang, J. Chun, *et al.*, "Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis," *Nature methods*, vol. 13, no. 3, p. 241, 2016.
16. C. Lin, S. Jain, H. Kim, and Z. Bar-Joseph, "Using neural networks for reducing the dimensions of single-cell rna-seq data," *Nucleic Acids Research*, 2017.
17. A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 1328–1335, IEEE, 2015.
18. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
19. J. M. Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*, pp. 720–722, Springer, 2011.
20. T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
21. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
22. I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*, pp. 115–128, Springer, 1986.
23. X. Yu, G. Yu, and J. Wang, "Clustering cancer gene expression data by projective clustering ensemble," *PloS one*, vol. 12, no. 2, p. e0171429, 2017.
24. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, *et al.*, "Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.
25. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq," *Science*, vol. 352, no. 6282, pp. 189–196, 2016.

26. H. Ishiura, W. Sako, M. Yoshida, T. Kawarai, O. Tanabe, J. Goto, Y. Takahashi, H. Date, J. Mitsui, B. Ahsan, *et al.*, "The trk-fused gene is mutated in hereditary motor and sensory neuropathy with proximal dominant involvement," *The American Journal of Human Genetics*, vol. 91, no. 2, pp. 320–329, 2012.
27. K. Kikuta, D. Kubota, T. Saito, H. Orita, A. Yoshida, H. Tsuda, Y. Suehara, H. Katai, Y. Shimada, Y. Toyama, *et al.*, "Clinical proteomics identified atp-dependent rna helicase ddx39 as a novel biomarker to predict poor prognosis of patients with gastrointestinal stromal tumor," *Journal of proteomics*, vol. 75, no. 4, pp. 1089–1098, 2012.
28. I. Redonnet-Vernhet, D. J. Mahuran, R. Salvayre, F. Dubas, and T. Levade, "Significance of two point mutations present in each hexb allele of patients with adult gm2 gangliosidosis (sandhoff disease) homozygosity for the ile207 val substitution is not associated with a clinical or biochemical phenotype," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1317, no. 2, pp. 127–133, 1996.
29. R. Nohra, A. Beyeen, J. Guo, M. Khademi, E. Sundqvist, M. Hedreul, F. Sellebjerg, C. Smestad, A. Oturai, H. Harbo, *et al.*, "Rgma and il21r show association with experimental inflammation and multiple sclerosis," *Genes and immunity*, vol. 11, no. 4, pp. 279–293, 2010.
30. J. H. Zar, "Spearman rank correlation," *Encyclopedia of Biostatistics*, 1998.
31. F. Arcuri, S. Papa, A. Carducci, R. Romagnoli, S. Liberatori, M. G. Riparbelli, J.-C. Sanchez, P. Tosi, and M. T. del Vecchio, "Translationally controlled tumor protein (tctp) in the human prostate and prostate cancer cells: expression, distribution, and calcium binding activity," *The Prostate*, vol. 60, no. 2, pp. 130–140, 2004.
32. Y. Dai, S. E. Pierson, W. Dudney, and B. C. Stack, "Extraribosomal function of metallopanstimulin-1: reducing paxillin in head and neck squamous cell carcinoma and inhibiting tumor growth," *International journal of cancer*, vol. 126, no. 3, pp. 611–619, 2010.
33. A. Geirsson, R. J. Lynch, I. Paliwal, A. L. Bothwell, and G. L. Hammond, "Human trophoblast noncoding rna suppresses ciita promoter iii activity in murine b-lymphocytes," *Biochemical and biophysical research communications*, vol. 301, no. 3, pp. 718–724, 2003.
34. E. Vairaktaris, Z. Serefoglou, D. Avgoustidis, C. Yapijakis, E. Critselis, A. Vylliotis, S. Spyridonidou, S. Derka, S. Vassiliou, E. Nkenke, *et al.*, "Gene polymorphisms related to angiogenesis, inflammation and thrombosis that influence risk for oral cancer," *Oral oncology*, vol. 45, no. 3, pp. 247–253, 2009.
35. T. Funasaka, A. Haga, A. Raz, and H. Nagase, "Tumor autocrine motility factor is an angiogenic factor that stimulates endothelial cell motility," *Biochemical and biophysical research communications*, vol. 284, no. 5, pp. 1116–1125, 2001.
36. M. Abu-Odeh, T. Bar-Mag, H. Huang, T. Kim, Z. Salah, S. K. Abdeen, M. Sudol, D. Reichmann, S. Sidhu, P. M. Kim, *et al.*, "Characterizing ww domain interactions of tumor suppressor wwox reveals its association with multiprotein networks," *Journal of Biological Chemistry*, vol. 289, no. 13, pp. 8865–8880, 2014.
37. K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, "Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data," *Genome research*, vol. 17, no. 11, pp. 1665–1674, 2007.
38. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.
39. M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.
40. F. Chollet, "keras," *GitHub repository*, 2015.
41. L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

1 Appendix

1.1 Software

We have developed a python package for the Dhaka variational autoencoder using the Keras module [40]. The package is released as open source (`https://github.com/MicrosoftGenomics/Dhaka`). Since this is a probabilistic encoding of the genomic data, often we need to do multiple warm starts of the encoder to select the best encoding. For example, if we are interested in identifying clusters, from each projected encoding we will compute the silhouette score, and select the encoding that maximizes the score [41]. We have used multiple warm starts only for the synthetic data analysis. We did not use multiple warm starts for the copy number and gene expression data. The number of warm starts is a user parameter for the package (5 in case of synthetic dataset). The Dhaka package can also perform gene selection, if needed. We have three options for selecting informative genes for analysis.

- Coefficient of variation (CV) score: CV of gene i with expression profile $g_i \in R^{1\times m}$ is defined as $CV_i = std(g_i)/mean(g_i)$. Here m is the total number of cells.
- Entropy En: $En_i = -sum(p_i log2(p_i))$. Here p_i is the estimated histogram from g_i.
- Average expression value $\bar{A}$: This is simply the average expression value of a particular gene across all cells.

The gene selection criteria and number of genes to be included in the analysis are both user parameters. We have used gene selection for the three RNA-Seq gene expression datasets, (5000 genes with $\bar{A}$ criteria). The variational autoencoder is robust to the drop out events, therefore we did not have to model the drop out events separately. The other parameters of the package are the number of the latent dimensions, learning rate, batch size, number of epochs, and clip norm of the gradient [1].

1.2 Relative gene expression

The relative gene expression $Er_{i,j} = E_{i,j} - mean(E_{i,1,\ldots,n})$. Here i and j correspond to gene and cell, respectively.

1.3 Robustness to drop out genes in synthetic data

As we know, drop outs are a major issue in single cell RNA-seq data analysis. Hence we also tried to analyze the robustness of the Dhaka method to drop out genes. We have simulated the drop out effect by randomly picking a percentage of genes for each cell to be dropped out. The set of genes dropped out in each cell is not necessarily same since these are picked randomly. Fig. 11 shows the performance of the Dhaka method in terms of ARI as we gradually increase the percentage of drop out genes in each sample. We can see that autoencoder is the most robust when compared to t-SNE and PCA. Both t-SNE and PCA fail to detect more than one cluster when more than 5% drop out genes were introduced in the sample. The dataset we used here is the synthetic dataset with 2500 noisy genes out of 3000 genes.

[1] Gradients will be clipped when their L2 norm exceeds this value. This parameter is used for the stability of the gradient descent algorithm.

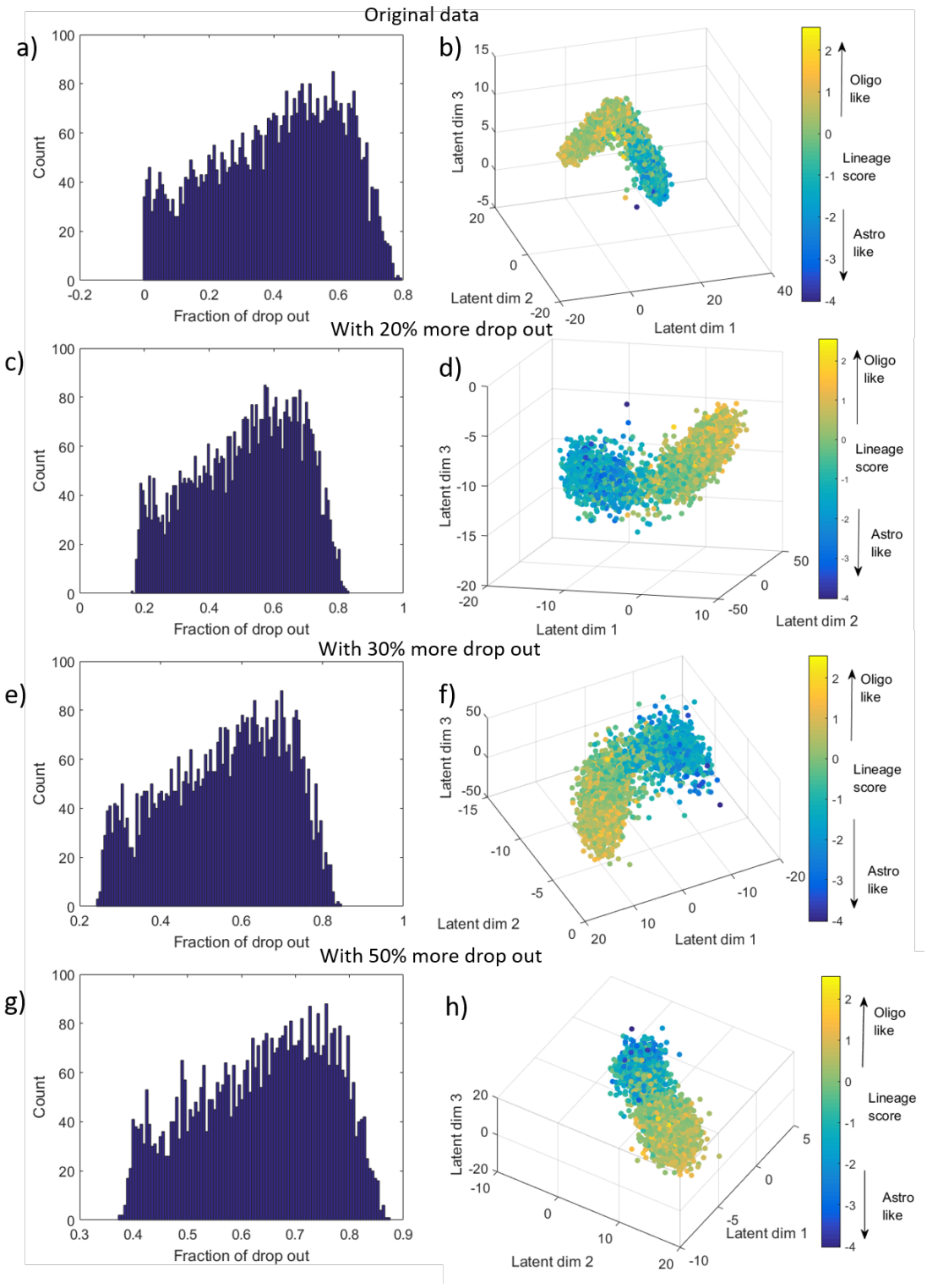

Fig. 10. Robustness analysis with Oligodendroglioma. a,c,e,g Histogram of drop out fraction in each gene after forcing 0%, 20%, 30%, and 50% more genes to be dropped out. b,d,f,h corresponding autoencoder projection of the data. We can see that upto 30%, the autoencoder can correctly identify v-structure. Beyond that the autoencoder loses the v-structure but still shows good sepration between oligo-like and astro-like cells.

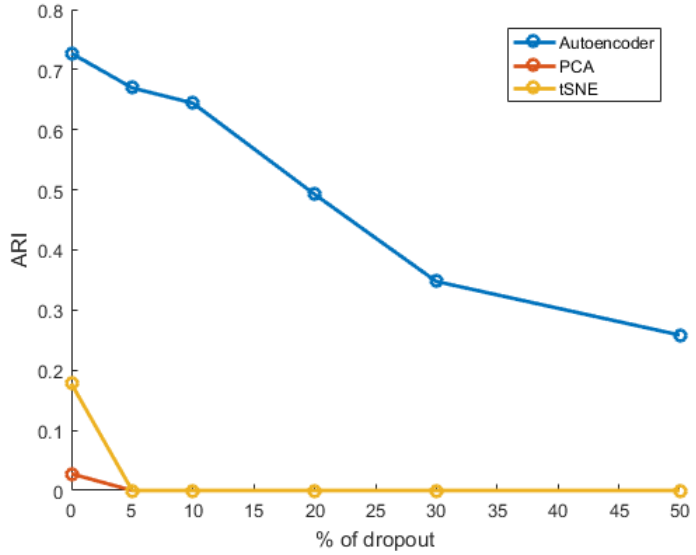

Fig. 11. Robustness to drop out genes in single cell expression data.

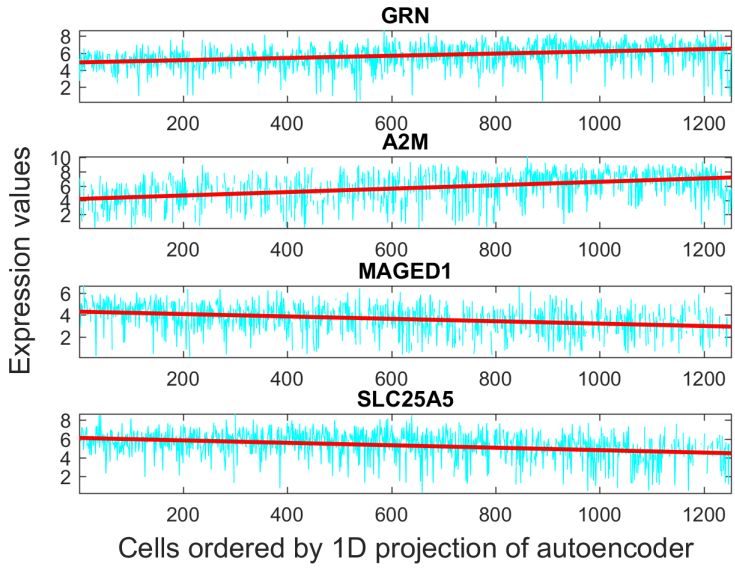

Fig. 12. Known marker genes for Melanoma MITF-AXL program.

1.4 Analysis of Astrocytoma data

The Astrocytoma dataset contains a total of 6341 cells with about $23K$ genes, among which 5097 are malignant cells. Astrocytoma is another type of brain tumor and this is a followup dataset from the Oligodendroglioma. Hence we performed same analysis as for Oligodendroglioma. The non-malignant microglia/macrophage cells were clearly separated from the malignant cells (Fig. 13a) in this dataset too. The authors did not compute differentiation and lineage metric for this dataset, but did mention that most of the cells fall in the intermediate state. When we fed the expression profile of the malignant cells to the autoencoder, it correctly placed most of the cells near the bifurcation point of the v-structure (Fig. 13b). For reference, we have also showed the Oligodendroglioma cells in the same plot colored by their differentiation score.

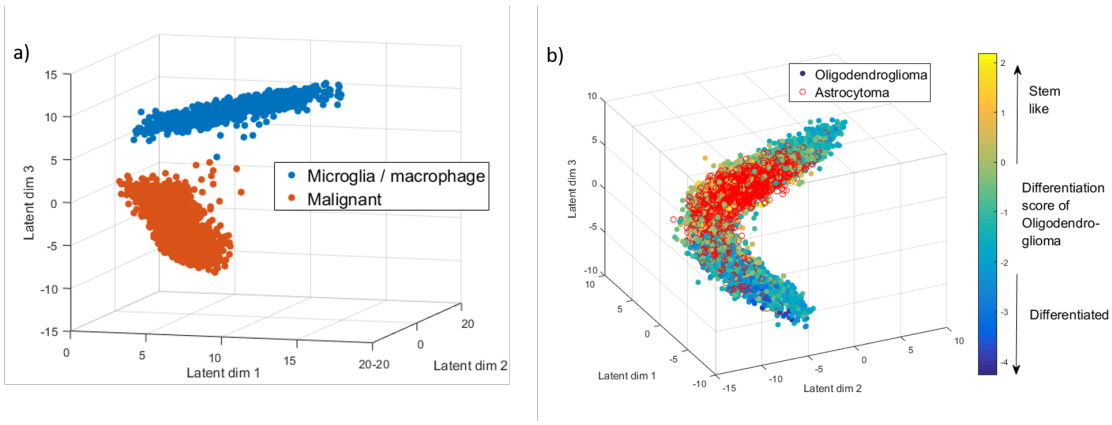

Fig. 13. Astrocytoma dataset. a) Autoencoder output of Astrocytoma dataset with 5000 autoselected genes separating malignant cells from microglia/macrophage cells. b) Autoencoder output from relative expression profile of malignant Astrocytoma cells(red) along with malignant Oligodendroglioma cells.