Carnegie Mellon University

Generative Adversarial Networks for scRNA-Seq Analysis

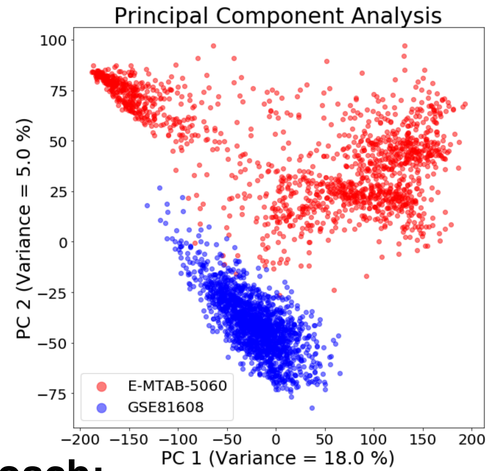
Akash Ramachandran, Donald Dalton

Summary:

In this study we assessed the performance of several unsupervised learning methods at the task of dimensionality reduction for single cell RNA-Seq data (scRNA-Seq). In addition to traditional dimensionality reduction methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), we also implemented a relatively new framework based on Generative Adversarial Networks (GANs).

Motivation:

For the last decade the cost of next generation sequencing (NGS) has decreased at a rate outpacing Moore's Law [1]. As such there is an increasing amount of biological data available for large scale genomic studies. However, integration of diverse biological datasets is a notoriously challenging problem primarily due to the sensitivity of NGS platforms to variable experimental conditions and protocols. Often it is the case that comparisons between datasets from multiple sources is not possible without sophisticated batch correction techniques which are often case-specific, and thus do not generalize well out of the box [2]. Therefore there is a need for frameworks that can effectively integrate diverse datasets with minimal pre-processing, and extract biologically relevant features in the presence of batch effects.



Approach:

Previous studies have shown that GANs can be used to learn a meaningful latent representation of diverse datasets [3]. Furthermore, the generative nature of these models can capture complex non-linear relationships amongst input variables and provides a flexible framework for post-hoc interpretation and analysis of learned features. Using the Wasserstein GAN [3] we performed dimensionality reduction of an aggregated scRNA-Seq dataset of human pancreatic cells.

Data:

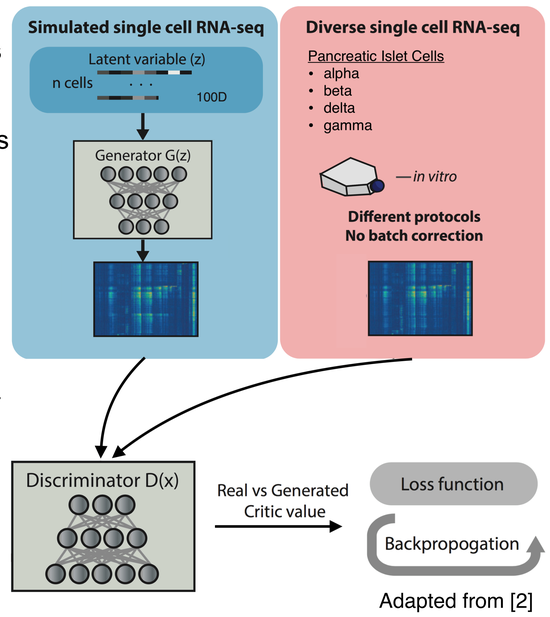
- 2 scRNA-Seq datasets
- 2 NGS Platforms
- 4 Cell Types
- 3K Samples, 8K Genes

Processing:

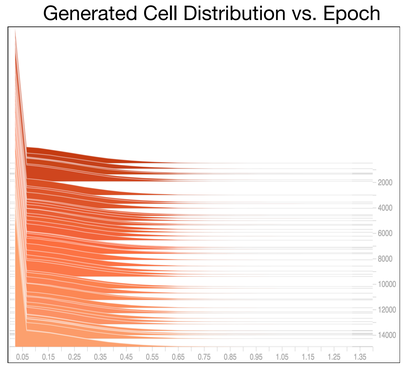
- RPKM Normalization
- Threshold Filtering
- Gene Intersection

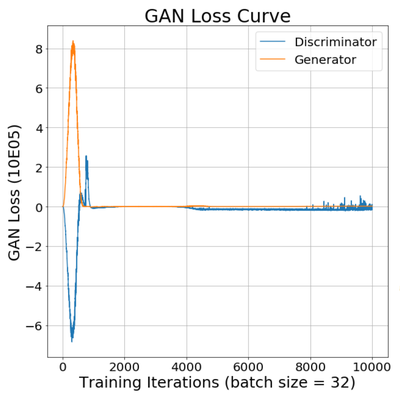
Model:

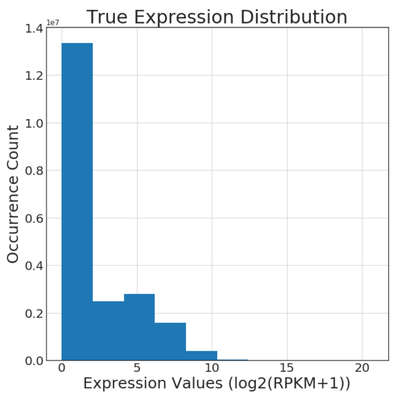
- **Generator:**
  - Input: 100 Unit Noise Vector
  - 600 Unit Hidden Layers (2)
  - Leaky ReLU Activation
  - Output: Fake Vectors
- **Discriminator**:
  - Input: Real+Fake Vectors
  - 100 Unit Hidden Layers (2)
  - Leaky ReLU Activation
  - Output: {Fake: 0, Real: 1}

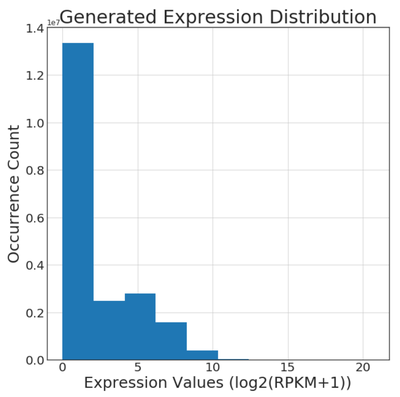


Results:

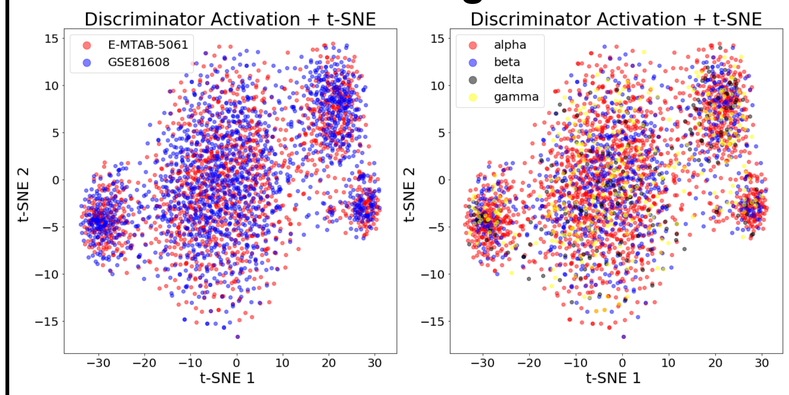




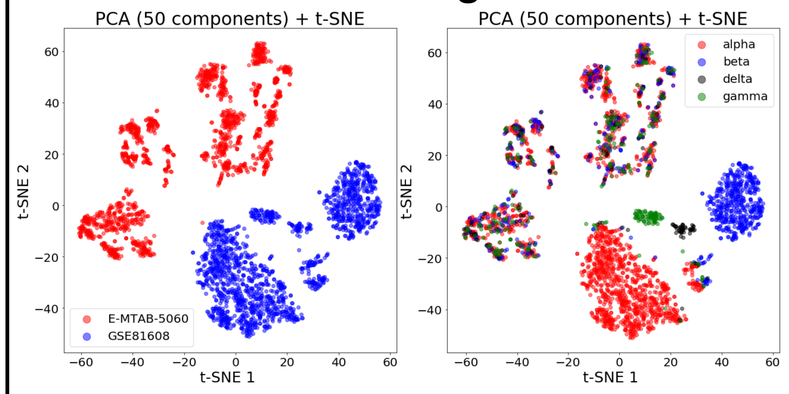




GAN+t-SNE Embeddings:



PCA+t-SNE Embeddings:



Classification:

KNN classification by cell type. 10 permutations, 10-fold cross-validation.

Algorithm	Precision	Recall	F1-Score	Support
PCA + t-SNE	**0.67 ±0.05**	**0.75 ±0.03**	**0.69 ±0.04**	307
GAN + t-SNE	0.52 ± 0.04	0.6 ± 0.03	0.49 ± 0.03	307

Future Work:

- Curation of a larger dataset
- Extract biological insights from latent representation

References:

1) DNA Sequencing Costs: Data [Internet] Bethesda (MD): National Human Genome Research Institute; 2016. [cited 2016 Oct 1].
2) Generative adversarial networks uncover epidermal regulators and predict single cell perturbations A Ghahramani, FM Watt, NM Luscombe - bioRxiv, 2018
3) M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. arXiv preprint arXiv: 1701.07875, 2017.