**DeepCyTOF**

<https://www.ncbi.nlm.nih.gov/pubmed/29036374>

<https://github.com/hzc363/DeepLearningCyTOF>

**DeepCyTOF** is to assign individual cells to known cell types. To accomplish this, DeepCyTOF methods first group and calibrate the cells into different clusters to both reference source samples and target samples, then find marker genes, and finally use these to manually assign target cell types for each cluster and identify its cell types. **DeepCyTOF** achieved the classification of cell types trained on a reference source sample automatically. However, it cannot detect new cells. At the same time, there is no comparison for embedding measurement from high-dimension to low-dimension.

(1) We use a denoising autoencoder (DAE) to handle missing data.

A DAE is a neural net that is trained to reconstruct a clean input from its corrupted version. Unlike (Vincent et al., 2010), who use Gaussian noise to corrupt the inputs, we use dropout noise, i.e., we randomly zero out subset of the entries of each cell, to simulate the machine instabilities. We train a DAE for each batch, by combining all samples from that batch, selecting the cells with no zeros and using them as training set. For each DAE, we set the drop-out probability to be the proportion of zeros in the measurement of the corresponding batch. Once a DAE is trained, we pass all samples from its batch through it to denoise the data.

(2) DeepCyTOF used multiple distribution-matching residual networks (MMD-ResNets)(Shaham et al., 2016) to calibrate an arbitrary number of source. It calibrate each batch to a reference using the MMD-ResNet approach.

Instrument calibration causes variation across samples, such a situation is often referred to ‘batch effect’. In order to avoid gating each dataset separately (which therefore requires labeled sam-ples from each dataset), a domain adaptation procedure is used. Domain Adaptation is a set of techniques that allow the use of a learning scheme (or model) trained on data from a source domain with a given distribution, which can then be applied to a target do-main with a related but not equivalent distribution. The objective of domain adaptation is to minimize the generalization error of instances from the target domain (Daume´, 2009; Daume and Marcu, 2006).

MMD is a measure for a distance between distributions, which had been shown to be suitable for training of neural nets-based generative models (Dziugaite et al., 2015; Li et al., 2015). If is a reproducing kernel Hilbert space with a (universal) kernel function , the (squared) MMD between distributions over a space is defined as

*where* and are independent, and so areand

For calibration purposes, we want to find a map that brings the distribution of the source sample close to that of the target sample; we further assume that this map should be close to the identity.

Using *ResNets* consisting of three blocks, which are trained to minimizing the loss

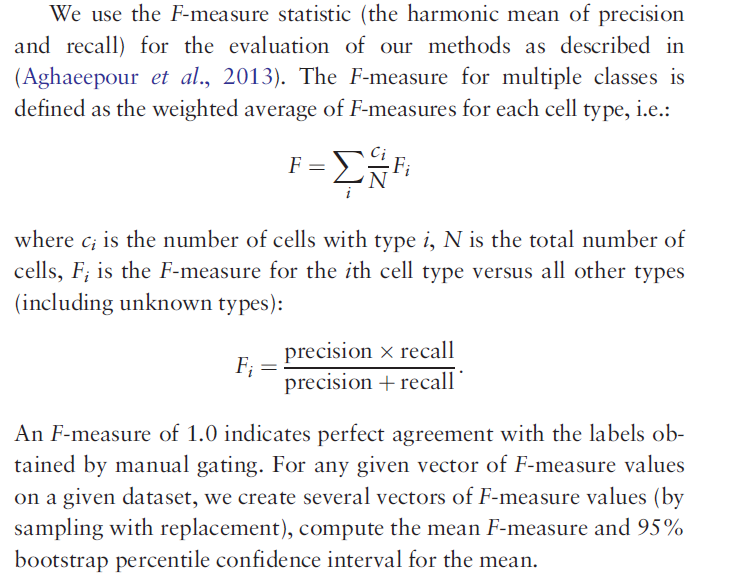
*such that*

(3) We performed a domain adaptation procedure for automatic classification.

**Datasets**:

We demonstrate the efficacy of DeepCyTOF in supplanting manual gating by first applying it to three CyTOF datasets consisting of 56, 136 and 16 PBMC samples respectively, and then comparing the concordance of the resultant cell classifications with those obtained by manual gating. Additionally, we benchmark DeepCyTOF’s pre-processing options for batch calibration using a collection of 16 bio-logical replicates measured in duplicates on eight CyTOF instruments. Finally, we compare DeepCyTOF to the other compet-ing supervised approaches benchmarked on each dataset of the forth challenge of the FlowCAP-I competition (Aghaeepour et al., 2013).

**Measurement:**



**[2] Automated identification of Cell Types in Single Cell RNA Sequencing**

[Ma F](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ma%20F%5BAuthor%5D&cauthor=true&cauthor_uid=31359028), [Pellegrini M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pellegrini%20M%5BAuthor%5D&cauthor=true&cauthor_uid=31359028). ACTINN: automated identification of cell types in single cell RNA sequencing. [*Bioinformatics*.](https://www.ncbi.nlm.nih.gov/pubmed/31359028) 2020 Jan 15;36(2):533-538. doi: 10.1093/bioinformatics/btz592.

**ACTINN** (Automated Cell Type Identification using Neural Networks) automatically learns the features for each predefined cell type and uses these features to predict cell types for individual cells. ACTINN allows users to rapidly identify cell types in their datasets, thus rendering the analysis of their Single cell RNA sequencing (scRNA-seq) datasets more efficient. This approach is computationally efficient and requires no domain expertise of the tissues being studied.

**Motivation:**

One common goal of scRNA-seq analyses is to identify the cell type of each individual cell that has been profiled. To accomplish this, typically cells are first grouped into different clusters in an unsupervised way, and the number of clusters allows us to approximately determine how many distinct cell types are present in the sample. Each cluster should contain cells with similar expression profiles, and so the aggregated profile of a cluster increases the signal to noise of the expression estimates. To attempt to interpret the identity of each cluster, marker genes are found as those that are uniquely highly expressed in a cluster, compared to all the other clusters. These canonical markers are then used to assign the cell types for the clusters, by cross referencing the markers with lists of previously characterized cell type specific markers. While this process is able to identify cell types, there are some limitations: 1. Since the clustering method is unsupervised, all sources of variation influence the formation clusters, including effects that are not directly related to cell types such as differential expression induced by cell cycles. 2. It is often difficult to find an optimal match between the marker genes associated with each cluster and the canonical markers for specific cell types. Moreover, depending on the clustering parameters used, one cluster might contain multiple cell types, or one cell type could be split into multiple clusters. 3. Using canonical markers to assign cell types requires background knowledge of cell type specific markers, and sometimes these are not well characterized or difficult to find in the literature. Moreover, some canonical markers may be expressed by more than one cell type, and some cell types may have no known markers. 4. The same types of cells processed by two distinct scRNA-seq techniques tend to cluster separately due to technical batch effects, which complicates cell type identification in composite datasets. 5. **Cell subtypes are often very similar to each other**, which limits efforts to separate them accurately into different clusters. To overcome many of the limitations of existing approaches, new methods need to be developed.

**Neural networks provide a popular framework for machine learning algorithms which can be used to interpret complex datasets. As a result, neural networks have been widely used in many fields, including for the analysis of scRNA-seq data [2-5]. Since the output data from scRNA-seq is feature-enriched and well-structured, it is well suited as an input for neural networks. Here, we present ACTINN (Automated Cell Type Identification using Neural Networks) for scRNA-seq cell type identification. To overcome may of the limitations of traditional cell type identification**

[2] Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res. 2017;45(17):e156.

[3] Shaham U, Stanton KP, Zhao J, et al. Removal of batch effects using distribution-matching residual networks. Bioinformatics. 2017;33(16):2539-2546.

[4] Lopez, R., et al., Deep generative modeling for single-cell transcriptomics. Nature Methods, 2018. 15(12): p. 1053-1058.

[5] Cho, H., Berger, B. & Peng, J. Generalizable and Scalable Visualization of Single Cell Data Using Neural Networks. Cell Syst 7, 185-191 e184, doi:10.1016/j.cels.2018.05.017 (2018).

Tamim Abdelaal, Vincent van Unen, Thomas Höllt, Frits Koning, Marcel J.T. Reinders, Ahmed Mahfouz. Predicting Cell Populations in Single Cell Mass Cytometry Data. Cytometry Part A, 95A: 769–781, 2019.

## **DBSCAN**

More formally, we define a core sample as being a sample in the dataset such that there exist min\_samples other samples within a distance of eps, which are defined as *neighbors* of the core sample. This tells us that the core sample is in a dense area of the vector space. A cluster is a set of core samples that can be built by recursively taking a core sample, finding all of its neighbors that are core samples, finding all of *their* neighbors that are core samples, and so on. A cluster also has a set of non-core samples, which are samples that are neighbors of a core sample in the cluster but are not themselves core samples. Intuitively, these samples are on the fringes of a cluster.

**Build the minimum spanning tree**

<https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html>

HDBSCAN is a clustering algorithm developed by [Campello, Moulavi, and Sander](http://link.springer.com/chapter/10.1007%2F978-3-642-37456-2_14). It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters.