Bioinformatics, 33(21), 2017, 3423–3430
doi: 10.1093/bioinformatics/btx448
Advance Access Publication Date: 13 July 2017
Original Paper

Gene expression

Gating mass cytometry data by deep learning

Huamin Li[1,†], Uri Shaham[2,†], Kelly P. Stanton[3], Yi Yao[4], Ruth R. Montgomery[4] and Yuval Kluger[1,3,5,*]

[1]Applied Mathematics Program and [2]Department of Statistics, Yale University, 51 Prospect Street, New Haven, CT 06511, USA, [3]Department of Pathology and Yale Cancer Center and [4]Department of Internal Medicine, Yale School of Medicine, New Haven, CT 06520, USA and [5]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally to this work.

Associate Editor: Janet Kelso

Received on March 21, 2017; revised on June 13, 2017; editorial decision on July 3, 2017; accepted on July 7, 2017

Abstract

Motivation: Mass cytometry or CyTOF is an emerging technology for high-dimensional multiparameter single cell analysis that overcomes many limitations of fluorescence-based flow cytometry. New methods for analyzing CyTOF data attempt to improve automation, scalability, performance and interpretation of data generated in large studies. Assigning individual cells into discrete groups of cell types (gating) involves time-consuming sequential manual steps, untenable for larger studies.

Results: We introduce DeepCyTOF, a standardization approach for gating, based on deep learning techniques. DeepCyTOF requires labeled cells from only a single sample. It is based on domain adaptation principles and is a generalization of previous work that allows us to calibrate between a target distribution and a source distribution in an unsupervised manner. We show that DeepCyTOF is highly concordant (98%) with cell classification obtained by individual manual gating of each sample when applied to a collection of 16 biological replicates of primary immune blood cells, even when measured across several instruments. Further, DeepCyTOF achieves very high accuracy on the semi-automated gating challenge of the FlowCAP-I competition as well as two CyTOF datasets generated from primary immune blood cells: (i) 14 subjects with a history of infection with West Nile virus (WNV), (ii) 34 healthy subjects of different ages. We conclude that deep learning in general, and DeepCyTOF specifically, offers a powerful computational approach for semi-automated gating of CyTOF and flow cytometry data.

Availability and implementation: Our codes and data are publicly available at https://github.com/KlugerLab/deepcytof.git.

Contact: yuval.kluger@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Flow cytometry (FCM) is routinely used in cellular and clinical immunology. Current fluorescence-based FCM experiments provide up to 15 numeric parameters for each individual cell from blood samples in a high-throughput fashion. Mass cytometry (CyTOF) is an emergent technological development for high-dimensional multiparameter single cell analysis. By using heavy metal ions as labels and mass spectrometry as readout, many more markers (>40) can be simultaneously measured. CyTOF provides unprecedented multidimensional immune cell profiling and has recently been applied to characterizing peripheral blood cells, Natural Killer cells in viral infections, skin cells, cells in celiac disease, responding phenotypes in cancer and even holds the promise of examination of solid tumors (Bendall *et al.*, 2011; Horowitz *et al.*, 2013; Strauss-Albee *et al.*, 2015; Yao *et al.*, 2014; Han *et al.*, 2013; Irish and Doxie, 2014; Giesen *et al.*, 2014; Angelo *et al.*, 2014). Cellular characterization by FCM and CyTOF will improve our understanding of disease processes (Atkuri *et al.*, 2015).

Gating (assigning individual cells into discrete groups of cell types) is one of the important steps and a bottleneck of analyzing FCM and CyTOF data. The time it takes to manually analyze a cytometry experiment depends on the number of blood samples as well as the number of markers (Verschoor *et al.*, 2015). Specifically, gating is performed by drawing polygons in the plane of every two markers. This implies that the time required for gating is roughly quadratic in the number of markers. In addition, the manual procedure, combined with the increase in the number of markers, make this process prone to human errors. Technical variation naturally arises due to the variation between individual operators (Benoist and Hacohen, 2011). The subjectivity of manual gating introduces variability into the data and impacts reproducibility and comparability of results, particularly in multi-center studies (Maecker *et al.*, 2012). Thus the slow processing time and the inherent subjectivity of manual analysis should be considered as primary reasons for using computational assistance methods.

Recently, deep learning methods have achieved outstanding performance in various computational tasks, such as image analysis, natural language processing, and pattern recognition (Deng and Yu, 2014). These approaches have also been shown to be effective for extracting natural features from data in general settings (Bengio *et al.*, 2013; Schmidhuber, 2015). Moreover, recent efforts to use deep learning approaches in genomics and biomedical applications show their flexibility for handling complex problems (Cireşan *et al.*, 2013; Cruz-Roa *et al.*, 2014; Denas and Taylor, 2013; Fakoor *et al.*, 2013; Leung *et al.*, 2014). However, deep learning typically requires very large numbers of training instances and thus its utility for many genomic, proteomic and other biological applications is questionable. While in most genomics applications, the number of instances (e.g. number of gene expression arrays) is typically smaller than the number of variables (e.g. genes), in each FCM and CyTOF run we typically collect approximately 10^5 to 10^6 cells, so that the number of instances (cells) is several orders of magnitude larger than the number of variables (up to 50 markers). Therefore, developing deep learning approaches to analyze cytometry data is very promising.

Importantly, in FCM and CyTOF experiments, variation in both biological and technical sources can make automatic gating challenging. 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 samples 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 domain with a related but not equivalent distribution. The objective of domain adaptation is to minimize the generalization error of instances from the target domain (Daumé, 2009; Daume and Marcu, 2006).

We present DeepCyTOF, an integrated deep learning domain adaptation framework, which employs one manually gated reference sample and utilizes it for automated gating of the remaining samples in a study. We first include two preprocessing options to use a denoising autoencoder (DAE) to handle missing data and use multiple distribution-matching residual networks (MMD-ResNets) (Shaham *et al.*, 2016) to calibrate an arbitrary number of source samples to a fixed reference sample, and then perform a domain adaptation procedure for automatic gating.

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 preprocessing options for batch calibration using a collection of 16 biological replicates measured in duplicates on eight CyTOF instruments. Finally, we compare DeepCyTOF to the other competing supervised approaches benchmarked on each dataset of the forth challenge of the FlowCAP-I competition (Aghaeepour *et al.*, 2013).

2 Materials and methods

DeepCyTOF integrates between three different tasks needed to achieve automated gating of cells in multiple target samples (the usage of the terms ‘source’ and ‘target’ in this manuscript is opposite than in (Shaham *et al.*, 2016), in order to be consistent with the domain adaptation terminology) based on manual gating of a single reference source sample. The tasks include sample denoising, calibration between target samples and a single reference source sample and finally cell classification. We implement each of these tasks using the following three neural nets: (i) a denoising autoencoder (DAE) for handling missing data; (ii) an MMD-ResNet for calibrating between the target samples and a reference source sample; (iii) a depth-4 feed-forward neural net for classifying/gating cell types trained on a reference source sample. DeepCyTOF has options to run with or without denoising, and with or without calibration.

2.1 Removing zeros using denoising autoencoder

All samples in our Mass Cytometry dataset contain large proportions of zero values across different markers. This usually does not reflect biological phenomenon, but rather, occurs due to instabilities of the CyTOF instrument. To tackle this, we include an option in DeepCyTOF to remove the zeros by training a denoising autoencoder (DAE) (Vincent *et al.*, 2010) on the cells with no or very few zero values. 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 dropout 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.

2.2 MMD-ResNet

To account for machine based technical bias and variability, we include a preprocessing option in DeepCyTOF to calibrate each batch to a reference using the MMD-ResNet approach. MMD-ResNet (Shaham *et al.*, 2016) is a deep learning approach to learn a map that calibrates the distribution of a source sample to match that of a target sample. It is based on a residual net (ResNet) architecture (He *et al.*, 2016a, 2016b) and has Maximum Mean discrepancy (MMD) (Borgwardt *et al.*, 2006; Gretton *et al.*, 2012) as the loss function. ResNet is a highly successful deep networks architecture, which is based on the ability to learn functions which are close to the identity. 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 $\mathcal{F}$ is a reproducing kernel Hilbert space with a (universal) kernel function $k(\cdot,\cdot)$, the (squared) MMD between distributions p, q over a space $\mathcal{X}$ is defined as

$$\mathrm{MMD}^2(\mathcal{F},p,q) = \mathbb{E}_{x,x'\sim p}k(x,x') - 2\mathbb{E}_{x\sim p,y\ \sim q}k(x,y) + \mathbb{E}_{y,y'\sim q}k(y,y'),$$

where x and x' are independent, and so are y and y'.

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. In a previous work (Shaham *et al.*, 2016), we have shown that MMD-ResNets are successful in learning such maps, and used them for calibration of CyTOF data and single cell RNA sequencing data. We refer the reader to (Shaham *et al.*, 2016) for a more comprehensive description of MMD-ResNets.

In this manuscript, we follow the approach of (Shaham *et al.*, 2016) and use ResNets consisting of three blocks, which are trained to minimizing the loss

$$L(w) \quad = \sqrt{\mathrm{MMD}^2\left(\tilde{f}(X), Y\right)}$$

such that

$$\mathrm{MMD}^2\left(\tilde{f}(X), Y\right) = \frac{1}{n^2}\sum_{x_i,x_j\in X} k\left(\tilde{f}(x_i), \tilde{f}(x_j)\right) - \frac{2}{nm}\sum_{x_i\in X, y_j\in Y} k\left(\tilde{f}(x_i), y_j\right) + \frac{1}{m^2}\sum_{y_i,y_j\in Y} k(y_i,y_j)$$

where $\tilde{f}$ is the map computed by the network, w are the network parameters, and $X=\{x_1,\ldots,x_n\}, Y=\{y_1,\ldots,y_m\}$ are two finite samples from the source and target distributions, respectively. An example of a MMD-ResNet is shown in Figure 1.

2.3 Cell classifier

To choose which sample will be used as reference source sample, for each sample i we first compute the $d\times d$ covariance matrix Σ_i, where d is the number of markers (dimensionality) of these samples. For every two samples i, j we then compute the Frobenius norm of the difference between their covariance matrices $||\Sigma_i - \Sigma_j||_F$, and we select the sample with the smallest average distance to all other samples to be the reference sample. Once the reference sample is chosen, we use manual gating to label its cells; the gating is used as ground truth labels to train the classifier. This is the only label information DeepCyTOF requires, regardless of the total number of samples we want to gate.

To classify cell types, we trained depth-4 feed-forward neural nets, each consisting of three softplus hidden layers and a softmax output layer. Further technical details regarding the architecture and training are given in Section 3.6. An example of such classifier is shown in Figure 2.

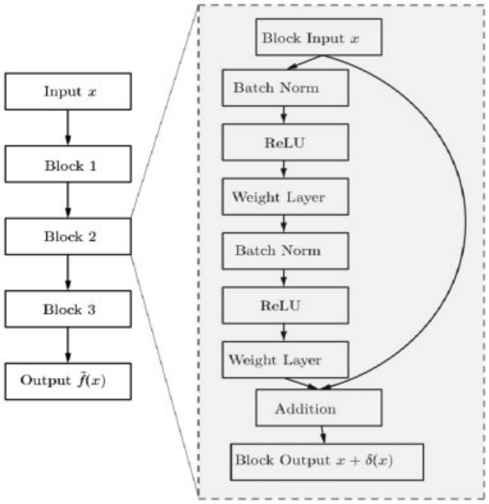

Fig. 1. MMD-ResNet with three blocks

3 Results

In this section, we present results from three experiments: (i) cell classification of five FCM datasets from the FlowCAP-I competition by applying DeepCyTOF without denoising (DAE) and without calibration (MMD-ResNets), as they are not needed as explained above, (ii) cell classification of two CyTOF datasets by applying DeepCyTOF with the denoising option (DAE), but without calibration (MMD-ResNets) of the target samples, (iii) cell classification of a multi-center CyTOF dataset by applying DeepCyTOF with the denoising option (DAE), and with calibration (MMD-ResNets) of the target samples to the source sample.

3.1 Datasets

3.1.1 FlowCAP-I datasets

We employ five collections of FCM datasets from FlowCAP-I (Aghaeepour *et al.*, 2013): (i) Diffuse large B-cell lymphoma (DLBCL), (ii) Symptomatic West Nile virus (WNV), (iii) Normal donors (ND), (iv) Hematopoietic stem cell transplant (HSCT) and (v) Graft-versus-host disease (GvHD). With the results from manual gates produced by expert analysis, the goal of FlowCAP-I challenges is to compare the results of assigning cell events to discrete cell populations using automated gates. In particular, we consider ‘Challenge 4: supervised approaches trained using human-provided gates’. We use the manual gating provided from FlowCAP-I to evaluate the neural nets predictions.

3.1.2 Mass cytometry datasets

We analyze two collections of CyTOF datasets measured on one instrument in the Montgomery Lab. The datasets consist of primary immune cells from blood of (i) $N=14$ subjects (8 asymptomatic and 6 severe) with a history of infection with West Nile virus (WNV),

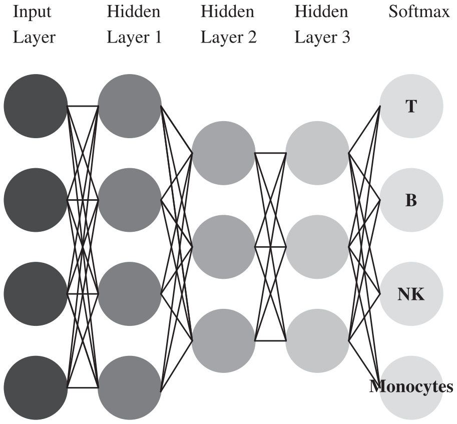

Fig. 2. A neural net for classifying cell types

and (ii) $N = 34$ healthy subjects of different ages (20 young and 14 old). Each blood sample is labeled with $d = 42$ antibody markers (Strauss-Albee *et al.*, 2015), 12 of which are used in our analysis as they are the relevant markers for the task of classification described below: HLA-DR, CD4, CD8, CD3-UCH1, CD16, CD33, CD19, CD14, CD56, DNA1, DNA2, Cisplatin. Each sample is subjected to four CyTOF experiments including a baseline state and three different stimuli (PMA/ionomycin, tumor cell line K562, and infection with WNV). The goal is to classify each cell to one of six cell type categories: (i) B cell, (ii) CD4+ T cell, (iii) CD8+ T cell, (iv) Monocytes, (v) Natural killer (NK) cells and (vi) unlabeled. There are 56 and 136 samples in the first two datasets, and we manually gate each single cell to evaluate the neural nets predictions.

We analyze an additional third collection of 16 CyTOF samples, which are all drawn from a single subject. These samples are measured in two different times and instruments as a part of a multi-center study (Nassar *et al.*, 2015). The first eight samples are collected at the same time, and the last eight samples are collected two months apart from the first eight. Additionally, each consecutive two samples are measured by the same instrument. Each of the 16 samples contains 26 markers, out of which eight correspond to the following protein markers: CCR6, CD20, CD45, CD14, CD16, CD8, CD3, CD4; we perform our classification experiments on this 8-dimensional data as these eight markers are the relevant ones for the task of classification: classify each cell to one of five cell type categories: (i) B cell, (ii) CD4+ T cell, (iii) CD8+ T cell, (iv) Monocytes and (v) unlabeled. We manually gate each single cell to evaluate the neural nets predictions.

3.1.3 Pre-processing

For FlowCAP-I datasets, we apply a logarithmic transform, followed by rescaling, as described in the Supplementary Material.

For Mass Cytometry datasets, we first manually filter all samples to remove debris and dead cells. In addition, different samples are measured at different times; fine changes in the state of the CyTOF instrument between these runs introduce additional variability into the measurements (batch effects). The specific nature of these changes is neither known nor modeled. To tackle this problem and apply a gating procedure, we follow most practitioners in the field, and calibrate the samples by applying an experimental-based normalization procedure (our results in Section 3.5 show that this normalization procedure does not always eliminate the batch effects between different instruments, and further calibration is needed). This procedure involves mixing samples with polystyrene beads embedded with metal lanthanides, followed by an algorithm which enables correction of both short-term and long-term signal fluctuations (Finck *et al.*, 2013). Once the data is normalized, we apply a logarithmic transform and rescaling.

3.2 Evaluation

To compare DeepCyTOF approach to algorithms from the forth challenge of the FlowCAP-I competition, 25% of the cells of each subject from the FCM datasets in FlowCAP-I are labeled by manual gating and used to train a cell type classifier based on the procedure of Section 2.3, which is then used to predict the labels of the remaining 75% cells. Here the DAE option is disabled because there are no missing values and the MMD-ResNet option is also disabled because the training and test sets are from the same run and thus do not require calibration.

To perform semi-automated gating of all samples of each of the first two CyTOF datasets based on the procedure of Section 2, we select a single baseline reference sample as in Section 3.1.2. We manually gate this sample, and use it to train a classifier for predicting the cell type class of each cell. The other baseline samples and additional samples that undergo three different stimuli (PMA/ionomycin, tumor cell line K562 and infection with WNV) are left for testing. Batch effects in these two datasets were not substantial allowing classification of cells into major cell populations without employing a calibration step as in Section 2.2 prior for to the cell classification of the test samples.

To perform semi-automated gating of the samples in the third multi-center CyTOF dataset, we follow the procedure of Section 2, i.e. we choose a single reference sample, train a collection of MMD-ResNets to calibrate all the remaining samples to it, train a cell type classifier on manually gated data of the reference sample, and use it to classify the cells of the calibrated samples. We compare the classification performance between two options of running DeepCyTOF. One option is with calibration of the target samples by MMD-ResNets and the other option is without calibration.

We use the *F*-measure statistic (the harmonic mean of precision and recall) for the evaluation of our methods as described in (Aghaeepour *et al.*, 2013). The *F*-measure for multiple classes is defined as the weighted average of *F*-measures for each cell type, i.e.:

$$F = \sum_i \frac{c_i}{N} F_i$$

where c_i is the number of cells with type i, N is the total number of cells, F_i is the *F*-measure for the ith cell type versus all other types (including unknown types):

$$F_i = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

An *F*-measure of 1.0 indicates perfect agreement with the labels obtained by manual gating. For any given vector of *F*-measure values on a given dataset, we create several vectors of *F*-measure values (by sampling with replacement), compute the mean *F*-measure and 95% bootstrap percentile confidence interval for the mean.

3.3 Evaluation of classification performance from FlowCAP-I

Table 1 presents the performance of DeepCyTOF when applied to the five datasets from the forth challenge of FlowCAP-I competition. The performance is compared to the respective winner of each of the five collections in this competition. As can be seen, our predictions are better than the competition winner in four out of the five collections and similar on the HSCT collection.

3.4 Application of DeepCyTOF to CyTOF datasets in the absence of strong batch effects

In this experiment, for each of the two different collections (which contain 14 and 34 baseline samples, respectively), we chose a

Table 1. Summary of results for the FlowCAP-I cell identification challenge

Dataset	DeepCyTOF	Competition's winner
GvHD	0.986 (0.979, 0.991)	0.92 (0.88, 0.95)
DLBCL	0.985 (0.976, 0.993)	0.95 (0.93, 0.97)
HSCT	0.991 (0.988, 0.993)	0.98 (0.96, 0.99)
WNV	0.999 (0.998, 0.999)	0.96 (0.94, 0.97)
ND	0.988 (0.987, 0.989)	0.94 (0.92, 0.95)

Note: The numbers in parentheses represent 95% confidence intervals for the *F*-measure. We use DeepCyTOF without denoising (DAE) and without calibration (MMD-ResNets).

Table 2. Summary of results for the two CyTOF collections

$F = \frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$	DeepCyTOF	Softmax regression
AnS-UN	0.990 (0.987, 0.993)	0.966 (0.955, 0.975)
AnS-N	0.991 (0.987, 0.994)	0.967 (0.957, 0.976)
OnY-UN	0.993 (0.990, 0.997)	0.963 (0.946, 0.977)
OnY-N	0.993 (0.989, 0.996)	0.963 (0.946, 0.977)

Note: The numbers in parentheses represent 95% confidence intervals. We use DeepCyTOF without usingMMD ResNets for calibration of the target samples. Datasets: unnormalized Asymptomatic&Severe (AnS-UN); normalized Asymptomatic&Severe (AnS-N); normalized 60 Old&Young (OnY N); unnormalized Old&Young (OnY-UN).

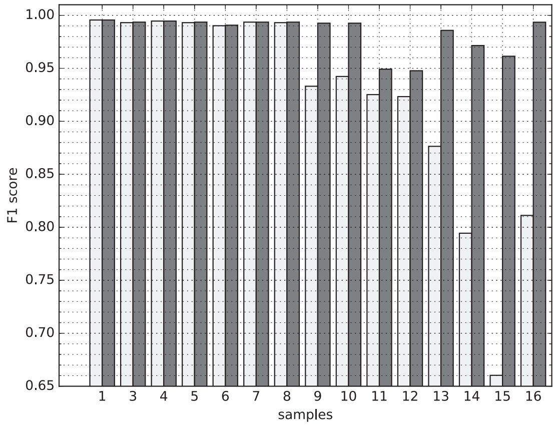

Fig. 3. Performance for the multi-center dataset before (white) and after (black) calibration

reference sample, used it to train DeepCyTOF using the options that omits the calibration step, which was then used to predict the cell types in all the other samples in that collection (55 samples from the Asymptomatic versus Severe WNV dataset and 135 samples from the Old versus Young dataset), without any calibration.

Supplementary Figure S1 shows an example of embedding of labeled cells in a three dimensional space, obtained from the top hidden layer of a neural net cell type classifier, as the ones used for this experiment. As can be seen, most of the labeled cells concentrate in separated clusters representing specific cell types. Table 2 summarizes the results, and provides a comparison to a shallow, linear classifier (softmax). Table 2 illustrates some interesting points: first, nearly perfect performance on the test data is achieved, despite the fact that it was only given labels from the reference sample. Second, DeepCyTOF performs significantly better than softmax regression, which may be a result of the depth and the non-linearity of the network. Third, whether or not the data is normalized does not affect the performance of DeepCyTOF. Fourth, we also applied DeepCyTOF choosing the option that includes a preprocessing calibration step in which MMD-ResNets are used to calibrate the source samples to the reference sample. However, this only yielded a modest improvement in the results. This may be due to the fact that the datasets did not significantly differ in distribution. A different scenario, where significant differences exist in the data, is considered in Section 3.5.

3.5 Overcoming strong batch effects using DeepCyTOF

The multi-center dataset consists of samples of the same subject, which were measured on different CyTOF machines in different locations. It is therefore reasonable that the different instrument conditions will result in calibration differences, which need to be accounted for. A domain adaptation framework like DeepCyTOF might be valuable in such scenario.

As in Section 2, we first chose a reference subject (sample 2) as the source. For each target sample, we then trained a MMD-ResNet to calibrate it to the reference sample. Subsequently, we trained a cell classifier using labeled data from the reference sample, and used it to classify cells from all the remaining calibrated 15 samples. We compared the performance of DeepCyTOF to a similar procedure where we skip the calibration step so that the input to the cell classifier is the data from the un-calibrated target samples. Figure 3 shows the *F*-measure scores for each sample before and after calibration. As can be seen, the *F*-measure scores of samples 9–16 are significantly higher when a calibration step is included in the gating process. For samples 1–8, we observed that the scores are very high even without a calibration step. Overall, applying DeepCyTOF with a calibration step results in an weighted average *F*-measure of 0.985 with 95% confidence interval (0.979, 0.990), which is significantly higher compared to the weighted average *F*-measure obtained by applying DeepCyTOF without calibration, which was 0.925 with 95% confidence interval (0.890, 0.956).

Figure 4 shows *t*-SNE embedding (Maaten and Hinton, 2008) of the cells of a representative sample selected from samples 1–8 (sample 7) and a representative sample from the other eight samples (sample 15) versus the cells of the reference sample, before and after calibration. On both samples the MMD-ResNet calibration seem to correct the batch effect appropriately, as after the calibration same cell types are embedded in the same clusters. To understand why the calibration almost did not change the accuracy on the last eight samples (while improving it dramatically on the last eight), Supplementary Figure S2 shows the MMD between the source sample and each of the target samples (the MMD values were computed using random batches of size 1000 from each of the distributions). As we can see, in all samples (with the exception of sample 2, which is the source sample, and sample 1 which was measured on the same instrument as sample 2), the MMD-ResNet calibration reduces the MMD between the distributions. However, before calibration the MMD between each the first eight target samples and the reference sample is relatively small, possibly making the classifier generalize well to these distributions even without calibration.

In the Supplementary Material, we also provide additional results from this experiment. A more detailed perspective on the effect of the calibration on the classification accuracy for the samples 9–16 is given in Figure 3, which shows the confusion matrix of a representative sample (sample 15), obtained before and after the calibration. For this sample, the *F*-measure obtained by applying DeepCyTOF with and without a calibration step are 0.9614 and 0.6603, respectively. In order to demonstrate the quality of calibration not only in a macroscopic level, but also when restricting the attention to a specific cell population, Supplementary Figure S4 shows the *t*-SNE plot of CD8 + T cells from sample 15 before and after calibration. The results are consistent with the ones given above. Finally, to demonstrate the effect of denoising on the quality of calibration, Supplementary Figure S5 shows the MMD between the reference sample and each of the other samples with and without denoising. We see that with denoising the MMD is smaller than without denoising.

3.6 Technical details

All cell type classifiers used in this work, were depth 4 feed-forward nets, with softplus hidden units and a softmax output layer, where the hidden layer sizes have been set to 12, 6 and 3. All MMD-ResNets were identical and consisted of three blocks, as can be seen

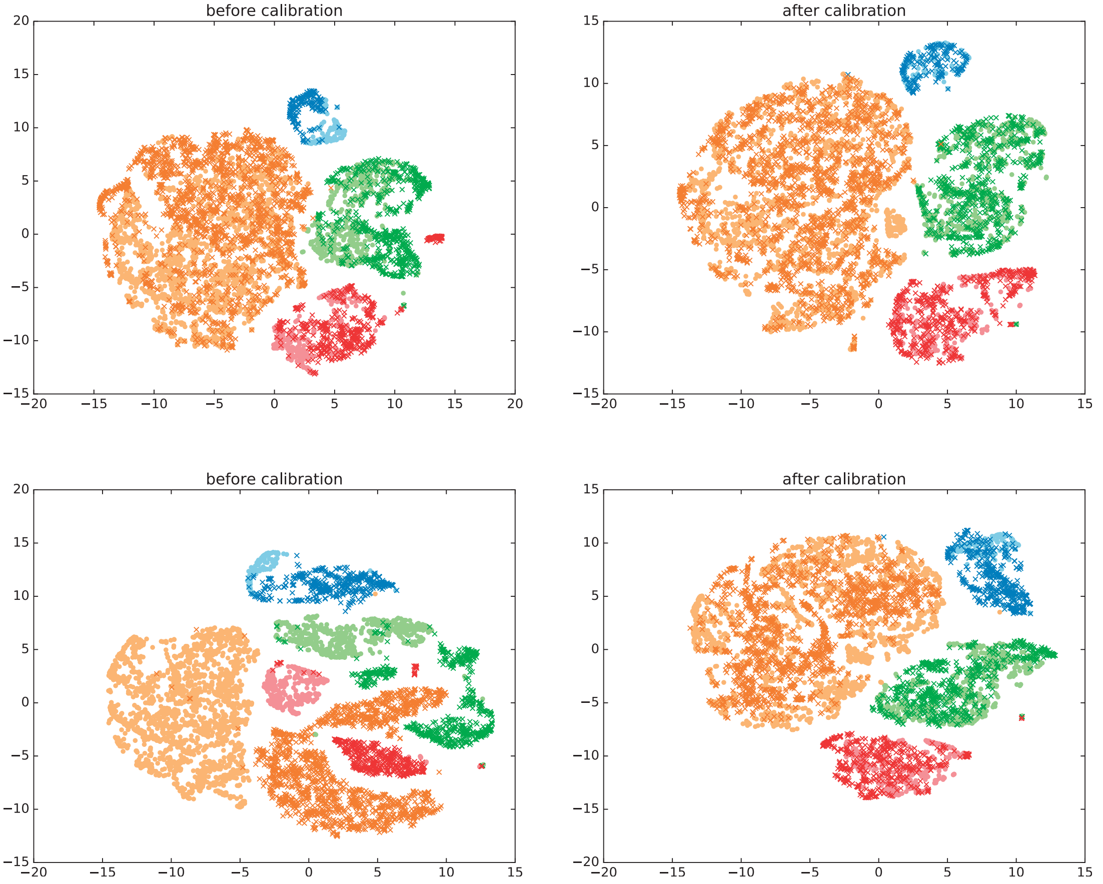

Fig. 4. Top: t-SNE plots of the joint distribution of sample 7 (dark crosses) and the reference sample (light circles) before (left) and after (right) calibration (the unlabeled cells are omitted). Bottom: Similarly to the upper panel but plots correspond to to the joint distribution of sample 15 and the reference sample. Different cell types have different colors: B cells (light and dark blue), CD4+ T cells (light and dark green), CD8+ T cells (light and dark red), Monocytes cells (light and dark orange). After calibration, same cell types are clustered together

in Figure 1. The first weight matrix was of size 25 × 8, and the second weight matrix was of size 8 × 25. The DAEs had two hidden layers, each of 25 ReLU units. All networks were trained using RMSprop (Tieleman and Hinton, 2012). We use the default learning rate (10^{-2}) to train the DAE. The learning rate schedule for training the classifiers and MMD-ResNets was defined as follows:

$$\mu(t) = \mu(0) \cdot \delta^{\lfloor \frac{1+t}{T} \rfloor}$$

where $\mu(t)$ was the learning rate at epoch t, $\mu(0)$ was the initial learning rate, δ was a constant, and T was the schedule. For training the classifiers, we have $\mu(0) = 10^{-3}$, $\delta = .5$ and $T = 50$. For training the MMD-ResNets, we have $\mu(0) = 10^{-3}$, $\delta = .1$ and $T = 15$.

We used mini-batches of size 128 for the cell type classifiers and the DAE, and 1000 for the MMD-ResNets. For each net, a subset 10% of the training data is held out for validation, to determine when to stop the training. In the DAE and cell type classifiers, a penalty of 10^{-4} on the l_2 norm of the network weights is added to the loss for regularization. In the MMD-ResNets we used for this purpose a penalty of 10^{-2}.

The kernel used for MMD-ResNets is a sum of three Gaussian kernels

$$k(x, y) = \sum_{i=1}^{3} \exp\left(-\frac{||x - y||^2}{\sigma_i}\right);$$

we set the the σ_is to be $\frac{m}{2}, m, 2m$, where m is the median of the average distance between a point in the target sample to its 25 nearest neighbors.

4 Theoretical justification for the calibration step of DeepCyTOF

In the classical domain adaptation setting (Ben-David *et al.*, 2010), a domain is a pair $(\mathcal{D}, f)$, where $\mathcal{D}$ is a distribution on an input space $\mathcal{X}$ and $f : \mathcal{X} \to \{0, 1\}$ is a labeling function. A hypothesis is a function $h : \mathcal{X} \to \{0, 1\}$. Given a domain $(\mathcal{D}, f)$, any hypothesis is associated with an error

$$\epsilon(h) = \mathbb{E}_{x \sim \mathcal{D}}[|h(x) - f(x)|].$$

Given a hypothesis, a source domain $(\mathcal{D}_s, f_s)$ and a target domain $(\mathcal{D}_t, f_t)$, one is interested in expressing the target error

$$\epsilon_t(h) = \mathbb{E}_{x \sim \mathcal{D}_t}[|h(x) - f_t(x)|]$$

in terms of the source error

$$\epsilon_s(h) = \mathbb{E}_{x \sim \mathcal{D}_s}[|h(x) - f_s(x)|].$$

This corresponds to expressing the error of a classifier, trained on the source data, on the target data.

Let $\mathcal{H}$ be hypothesis class of finite VC dimension. Ben David *et al.* (2010) prove that for every $h \in \mathcal{H}$, with a probability of at least $1 - \delta$,

$$\epsilon_t(h) \le \epsilon_s(h) + d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + C,$$

where

$$d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_s} [h(x) = 1] - \Pr_{x \sim \mathcal{D}_t} [h(x) = 1] \right|,$$

and C is a constant which does not depend on h.

By considering $\mathcal{H}$ to be the class of Parzen window classifiers, it can be shown (Sriperumbudur *et al.*, 2009; Long *et al.*, 2015) that

$$d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) \lesssim \mathrm{MMD}_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t).$$

This implies that calibrating the data by minimizing the MMD between the source and target distribution, the target error will get close to the source error. This is precisely the procedure DeepCyTOF performs.

5 Conclusions and future research

In this work, we show that deep learning machinery can be very effective in classification of cell types; the performance substantially surpasses the predictive accuracy of the methods presented in the forth challenge of the FlowCAP-I competition. In addition, we introduce DeepCyTOF, an automated framework for gating cell populations in cytometry samples. DeepCyTOF integrates deep learning and domain-adaption concepts. The labels obtained by manual gating of the reference sample are utilized in a domain-adaptation manner. These steps enable DeepCyTOF to inherently calibrate the major cell populations of multiple samples with respect to the corresponding cell populations of the reference sample. We analyze 208 CyTOF samples and observed nearly identical results to those obtained by manual gating (with mean F-measure ≥ 0.98).

In practice, run-to-run variations in CyTOF experiments both in the same instrument and between instruments are very common. These variations lead to significant batch effects in the datasets with samples collected at different runs. As a result, noticeable differences between the data distributions of the training data (manually gated reference sample) and the remaining unlabeled test data (the remaining samples) are observed, and an approach such as domain-adaptation is required to remove these biases. Bead-normalization is an approach introduced to mass cytometry as a pre-processing step to diminish the effect of such variations (Finck *et al.*, 2013). Interestingly, application of DeepCyTOF to unnormalized and bead-normalized data did not show an advantage of using the latter for the task of automated gating.

Flow cytometry and mass cytometry experiments involve data with dimensionality ranging between 10 and 40. Transforming the raw multivariate data to other representations may offer advantages for tasks such as automated gating or calibration. Finding good representations can be done either by manual investigation (hand crafting) or automated approaches. In recent years, deep learning approaches have been shown to be suitable for learning useful representations of data in the sense that they provide new sets of features that makes subsequent learning easier.

As cytometry analyses become widely used in research and clinical settings, automated solutions for analyzing the high dimensional datasets are urgently needed. Current practice in which samples are first subjected to manual gating are slowly substituted by automatic gating methods (Chester and Maecker, 2015). Major contributions to between-sample variations in cytometry experiments arise not only due to biological or medical differences but due to machine biases. Here, we demonstrate that a novel machine learning approach based on deep neural networks and domain adaptation can substitute manual gating as they both produce indistinguishable results. In future work, we will demonstrate that deep learning approaches could address other challenges in analyzing cytometry data. This includes tasks such as further development of unsupervised calibration of samples, and feature extraction for classification or visualization of multiple samples.

Acknowledgements

The authors would like to thank Catherine Blish for CyTOF reagents, Ronald Coifman and Roy Lederman for helpful discussions and suggestions. This research was funded by NIH grant 1R01HG008383-01A1 (Y.K.).

References

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Angelo,M. *et al.* (2014) Multiplexed ion beam imaging of human breast tumors. *Nat. Med.*, **20**, 436–442.

Atkuri,K.R. *et al.* (2015) Mass cytometry: a highly multiplexed single-cell technology for advancing drug development. *Drug Metab. Dispos.*, **43**, 227–233.

Ben-David,S. *et al.* (2010) A theory of learning from different domains. *Mach. Learn.*, **79**, 151–175.

Bendall,S.C. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–696.

Bengio,Y. *et al.* (2013) Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 1798–1828.

Benoist,C., and Hacohen,N. (2011) Flow cytometry, amped up. *Science*, **332**, 677–678.

Borgwardt,K.M. *et al.* (2006) Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, **22**, e49–e57.

Chester, C., and Maecker, H.T. (2015) Algorithmic tools for mining high-dimensional cytometry data. *J. Immunol.*, **195**, 773–779.

Cireşan,D.C. *et al.* (2013) Mitosis detection in breast cancer histology images with deep neural networks. In: Mori,K. *et al. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. MICCAI 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, vol. 8150, pp. 411–418.

Cruz-Roa,A. *et al.* (2014) Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Gurcan,M.N. and Madabhushi,A. (eds) *SPIE Medical Imaging*. International Society for Optics and Photonics, San Diego, California, vol. 9041, pp. 904103–904103.

Daumé,H.,III (2009) Frustratingly easy domain adaptation. *arXiv*. preprint arXiv:*0907.1815*.

Daume, H., III and Marcu,D. (2006) Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.*, **26**, 101–126.

Denas,O., and Taylor,J. (2013) Deep modeling of gene expression regulation in an erythropoiesis model. In: *Representation Learning*, *ICML Workshop*. Appeared in the 30th International Conference on Machine Learning workshop on Representation Learning, Atlanta, Georgia, USA.

Deng,L., and Yu,D. (2014) Deep learning: methods and applications. *Found. Trends Signal Process.*, **7**, 197–387.

Dziugaite,G.K. *et al.* (2015) Training generative neural networks via maximum mean discrepancy optimization. *arXiv. preprint arXiv:1505.03906*.

Fakoor,R. *et al.* (2013) Using deep learning to enhance cancer diagnosis and classification. In: *Proceedings of the 30th International Conference on Machine Learning,* Atlanta, Georgia, USA. JMLR: W&CP, vol. 28.

Finck,R. *et al.* (2013) Normalization of mass cytometry data with bead standards. *Cytometry A*, **83**, 483–494.

Giesen,C. *et al.* (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, **11**, 417–422.

Gretton,A. *et al.* (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.

Han,A. *et al.* (2013) Dietary gluten triggers concomitant activation of cd4+ and cd8+ $\alpha\beta$ t cells and $\gamma\delta$ t cells in celiac disease. *Proc. Natl. Acad. Sci. U S A.*, **110**, 13073–13078.

He,K. *et al.* (2016a) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

He,K. *et al.* (2016b) Identity mappings in deep residual networks. In: Leibe,B. *et al.* (eds) *Computer Vision – ECCV 2016*. ECCV 2016. Lecture Notes in Computer Science, Springer, Cham, vol. 9908.

Horowitz,A. *et al.* (2013) Genetic and environmental determinants of human nk cell diversity revealed by mass cytometry. *Sci. Transl. Med.*, **5**, 208ra145–208ra145.

Irish,J.M., and Doxie,D.B. (2014) High-dimensional single-cell cancer biology. In: Fienberg,H. and Nolan,G. (eds) *High-Dimensional Single Cell Analysis*. Current Topics in Microbiology and Immunology. Springer, Berlin, Heidelberg, vol. 377.

Leung,M.K. *et al.* (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics*, **30**, i121–i129.

Li,Y. *et al.* (2015) Generative moment matching networks. In: *ICML. Proceedings of the 32nd International Conference on Machine Learning*, Lille, France. JMLR: W&CP, vol. 37, pp. 1718–1727.

Long,M. *et al.* (2015) Learning transferable features with deep adaptation networks. In: *ICML. Proceedings of the 32nd International Conference on Machine Learning*, Lille, France. JMLR: W&CP, vol. 37, pp. 97–105.

Maecker,H.T. *et al.* (2012) Standardizing immunophenotyping for the human immunology project. *Nat. Rev. Immunol.*, **12**, 191–200.

Nassar,A. *et al.* (2015) The first multi-center comparative study using a novel technology mass cytometry time-of-flight mass spectrometer (cytof2) for high-speed acquisition of highly multi-parametric single cell data: a status report. In: *Presented at the 30th Congress of the International Society of Advancement of Cytometry*.

Schmidhuber,J. (2015) Deep learning in neural networks: an overview. *Neural Netw.*, **61**, 85–117.

Shaham,U. *et al.* (2016) Removal of batch effects using distribution-matching residual networks. doi:10.1093/bioinformatics/btx196.

Sriperumbudur,B.K. *et al.* (2009) Kernel choice and classifiability for rkhs embeddings of probability distributions. In: Bengio,Y. *et al.* (eds) *Advances in Neural Information Processing Systems 22*, pp. 1750–1758.

Strauss-Albee,D.M. *et al.* (2015) Human NK cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Sci. Transl. Med.*, **7**, 297ra115–297ra115.

Tieleman,T., and Hinton,G. (2012) *Lecture 6.5—RmsProp: divide the gradient by a running average of its recent magnitude*. In: COURSERA: Neural Networks for Machine Learning, **4**(2), pp. 26–31.

Van der Maaten,L., and Hinton,G. (2008) Visualizing data using t-sne. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Verschoor,C.P. *et al.* (2015) An introduction to automated flow cytometry gating tools and their implementation. *Front. Immunol.*, **6**, 380.

Vincent,P. *et al.* (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, **11**, 3371–3408.

Yao,Y. *et al.* (2014) Cytof supports efficient detection of immune cell subsets from small samples. *J. Immunol. Methods*, **415**, 1–5.