*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject

# Tumor Deconvolution from Bulk Gene Expression Data

**Noah Cohen Kalafut** [1,*]

[1] Computer Science, University of Wisconsin-Madison, Madison, 53715, United States of America

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** The relative infiltration of certain cells common to tumor tissue have an impact on effective patient care, treatment methods, and cancer behavior (DREAM (2019), Becht *et al.* (2016)). Single-cell technologies can be used in order to examine relative infiltration. However, single-cell technologies remain expensive while options for collecting bulk expression data are significantly cheaper. As such, it would be very useful to be able to determine relative stromal and immune infiltration of these cells using bulk expression data. In the current age, there are many public resources and repositories containing potential training data. This is the topic of the 'Tumor Deconvolution' DREAM challenge upon which this paper is based DREAM (2019). To examine the efficacy of Bisque, MuSiC, and single-source non-ensemble SCDC for tumor deconvolution, we test the best-case accuracy and robustness of the algorithms. In addition, auxillary tests are performed to limit potential confounding factors in the data.
**Results:** MuSiC was found to be the most robust method for use with tumor deconvolution, followed by Bisque and single-source non-ensemble SCDC. For best-case data, the three methods generally maintained the same accuracy.
**Contact:** nkalafut@wisc.edu

## 1 Introduction

Breast and colorectal cancers were among the 3 most commonly diagnosed cancers in 2020, affecting 4.3 million people diagnosed last year. The cancers were also ranked as 2 of the top 5 causes of cancer-related deaths (Sung *et al.* (2021)). Any aid in the treatment of these cancers could then have a large impact on general well-being of those at risk or undergoing treatment.

A correlation has been found between the behavior of certain types of cancer and various stromal and immune cells (Becht *et al.* (2016)). For this paper, we will consider cells of interest B-Cells, CD4 T-Cells, CD8 T-Cells, Monocytes, Neutrophils, and NK Cells. This would be referred to as a partial 'coarse-grained' immune and stromal infiltration on the original DREAM challenge (DREAM (2019)).

The general approach would be to utilize single-cell technologies to obtain these relative infiltrations. There exists a much cheaper option, however. Bulk expression data is cheaper to acquire and a great amount of testing data can be sourced online. Interpretation of bulk expression can be more difficult than single-cell, which is where our primary question lies.

We wish to determine the efficacy of existing methods for coarse-grained tumor deconvolution purely using bulk expression data. Among the methods to be tested are MuSiC, Bisque, and SCDC.

(Jew *et al.* (2020)) Bisque takes reference single-cell data to determine the composition of provided bulk expression data. There exists another implementation that utilizes marker genes rather than single-cell data, but we will focus on the former to retain similarity with our other methods. There could be potential issues in the underlying assumption of Bisque. Namely, Bisque assumes that the input single-cell and bulk expression data are samples of the same tissue. Since we are predicting infiltration solely based on bulk expression data, this will not be possible.

(Wang *et al.* (2019)) MuSiC, similar to Bisque, takes reference single-cell data to determine the relative infiltration of given cells for provided bulk expression data. Bisque has a potential issue in taking single-cell and bulk data from different sources. MuSiC attempts to remedy this issue by creating a weighting scheme that considers gene expression variance between subjects. Genes that vary across subjects are given lower weight while consistent, informative genes are given higher weight to be used in the eventual deconvolution of the provided bulk expression data.

(Dong *et al.* (2021)) The founding paper for SCDC was published in 2021. As input, the method takes single-cell reference data and bulk

expression data to predict. SCDC has a unique attribute in its ability to use an ensemble of various single-cell references to predict relative cell abundances. For this project, however, this and subject-based adjustment will not be utilized for the sake of comparison with other methods. It is claimed that, when compared against existing methods without the use of an ensemble, the algorithm is similar to MuSiC but is more reliable across read magnitudes. The algorithm also automates filtering of potentially misclassified single-cell readings.

## 2 Methods

In assessing performance, there is always an inherent issue of generality. A variety of datasets need to be tested on each proposed solution to obtain a proper comparison. Additionally, methods that utilize different input parameters need to be compared with that bias in mind.

### 2.1 Datasets

(Allantaz *et al.* (2012)) GSE28490 provides healthy subject microarray gene expression data for all of our target cells. The samples were taken in 2012 with the 'HG-U133 Plus 2.0'. 5 pools of 5 subjects each were sampled for each cell (excluding monocytes, for which 10 pools were used). The isolated samples were then profiled, producing a total of 47 average single-cell samples. This matches with a good amount of our testing data discussed later in the section. This will be a source for our single-cell data.

(Du *et al.* (2006)) GSE72642 was sampled in 2015 as a survey of major blood cell types with the aim of linking the differentially expressed genes to those associated to select disorders of the blood. Provided are 3 samples for B-Cells, CD4 T-Cells, CD8 T-Cells, Monocytes, and NK Cells from healthy donors. This will be a source for our single-cell data.

(DREAM (2019)) Our subject DREAM challenge provides 'leaderboard' datasets which consist of bulk expression data and known relative abundances of our desired cells. There are 20 datasets annotated by cancer type for a total of 836 known expression-infiltration pairs. The data was measured using a variety of platforms from Affymetrix and Illumina. The most common platform is the 'HG-U133 Plus 2.0' with 7 datasets, 4 (83 pairs) of which use the whole range of genes, 1 (20 pairs) of which has several important genes removed, and 2 (25+30 pairs) of which have around half of the genes removed. This will be our bulk gene expression data for testing.

### 2.2 Performance

We will look at the error rate across combinations of datasets, samples, and cells. We can first examine the error on our 4 testing samples with full reads. This will be conducted separately for multiple input single-cell datasets to determine the consistency of our references.

We can then examine the error rate on these datasets by cell type, noting any specific oddities. It is necessary to account for differing cell magnitudes and biases within the testing dataset during this analysis step. We can achieve this by providing a reference for cell magnitude across the chosen testing datasets. In this way, we can compare cells of similar magnitude to look for errors of interest in the methods being evaluated.

The aforementioned tests evaluate the best-case scenario. For testing robustness, we can artificially limit our read data. This has already been done for 3 of our testing samples, but we can look at the relationship between accuracy and gene percentage to assess the robustness of each algorithm. This might generally favor certain patterns when taking away random values. However, since we are only comparing between methods, this is fine.

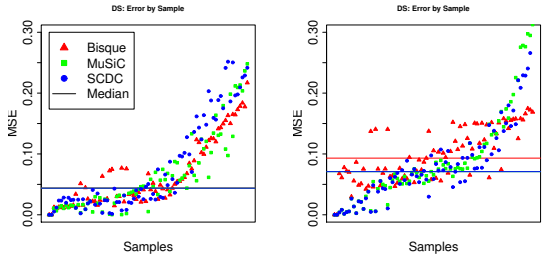



**Fig. 1.** GSE28490 (Left) compared to GSE72642 (Right) sorted by mean sample MSE

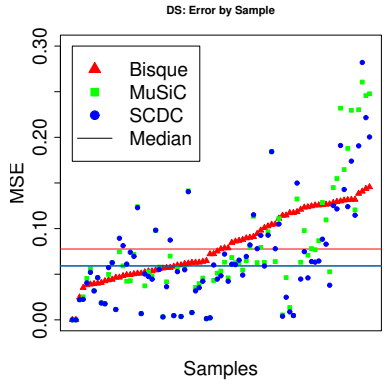



**Fig. 2.** GSE110085 (Left) compared to GSE13791 (Right) sorted by mean sample MSE

## 3 Results

### 3.1 Data Consistency

While comparing our datasets, the results for each dataset GSE28490 and GSE72642 remained similar in trend for all methods (figure 1). However, it appears that non-extreme values have increased in error per sample due to the higher middle samples and median. Bisque appears to have attained a large amount of noise in addition to this, but, after sorting by Bisque error rather than mean error, it can be seen that Bisque has actually lost accuracy in the same manner (figure 2). These results suggest that our algorithms prefer GSE28490 for tumor deconvolution on our dataset. More importantly, however, it points out a weakness of Bisque in the lack of standardization and single-cell quality control (such as cross-subject considerations) relative to MuSiC and SCDC.

The increase in error exceeds that caused by the exclusion of Neutrophils – which provides an adjusted MSE $>$ .06. This further suggests that GSE28490 is better suited for our tumor deconvolution. Moving forward, this dataset is used for the sake of simplicity.

### 3.2 General Performance

The three methods perform relatively equally in general. This makes sense as the three all have a similar base algorithm. The lack of increased accuracy for MuSiC and SCDC suggests suitable (or at least seemingly suitable) input single-cell data. The accuracies for each cell generally move in the direction of the average value for the testing dataset (figure 3).

### 3.3 Robustness

From figure 4, we are able to observe the flattening of the error curve and the increase in median MSE as data is taken away. The MSE is impacted, but the effect is generally equal for the three methods. Notice that SCDC

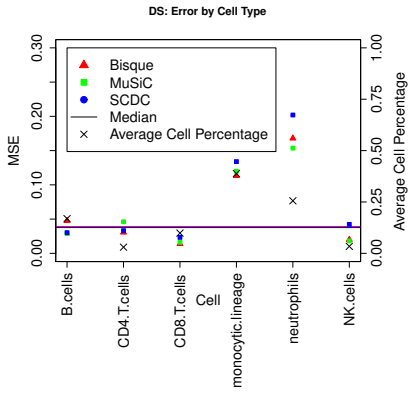



**Fig. 3.** Per-cell error on GSE28490 for full-read testing datasets

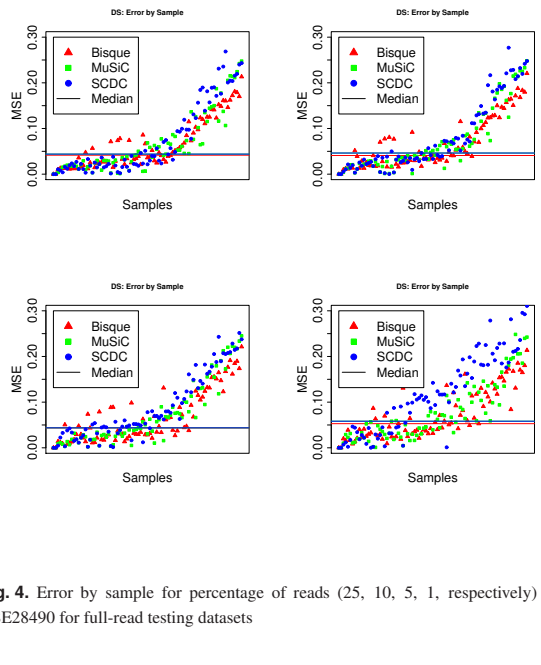



**Fig. 4.** Error by sample for percentage of reads (25, 10, 5, 1, respectively) used on GSE28490 for full-read testing datasets

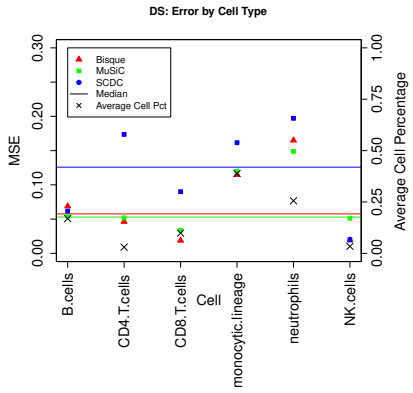
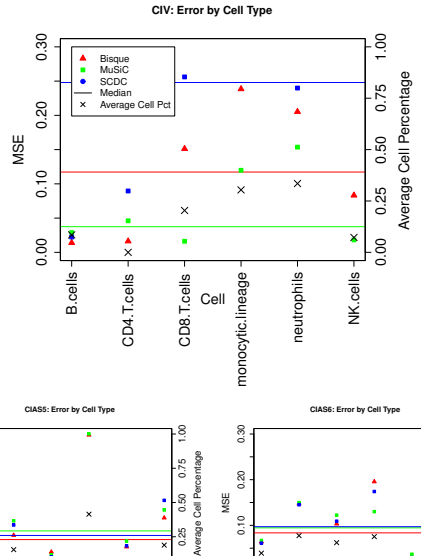



**Fig. 5.** Error by sample for 1 percent of reads kept used on GSE28490 for full-read testing datasets



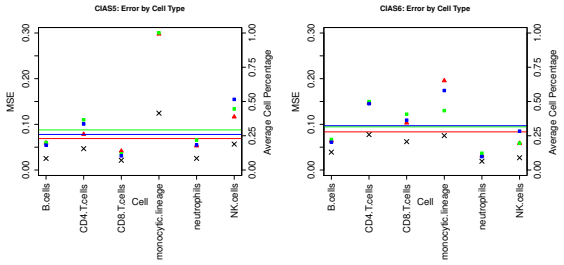




**Fig. 6.** Per-cell error on GSE28490 for select reads removed (CIVA1), half genes removed (CIAS5 & CIAS6)

## 4 Discussion

The two major outliers for our general data, given by figure 3, are CD4 T-Cells and Neutrophils. This can be explained by figure 7. Neutrophils are significantly differentially expressed from other cells in the dataset. While this may make Neutrophils more identifiable, it makes sense that a classifiers would mistake some number of Neutrophils for the next closest cell, Monocytes. The same goes for CD4+ and CD8+ T-Cells. The increased error on Neutrophils for SCDC is not surprising, given its more aggressive culling techniques.

In general, for apt input data, our three methods appear to be roughly equal (1). Bisque appears to be the most unstable. This is most likely due to the less advanced filtering and normalization techniques utilized.

In figure 5, we can more closely examine the effects of the data retention on the MSE for individual cells. As is evident, there has been a much more generally maintains a higher-error curve than the other two methods as the percentage of data retention lowers, despite having only a slightly higher median in the worst case. In figure 5, we observe this pattern in the extreme case.

dramatic effect on the accuracy of the individual cell reads as percentage genes retained approaches 1 percent. This in combination with figure 6, which demonstrates the same principle with more carefully selected genes, allows us to create a rough ordering of our methods.

As we make the dataset less reliable, SCDC became significantly less reliable than both Bisque and MuSiC, with MuSiC being the stronger of the two. This effect is most exemplified by figure 5.

Summarizing, we can say that single-source non-ensemble SCDC performs worse than both Bisque and MuSiC. Under these conditions, MuSiC is also the most robust in the worst-case.

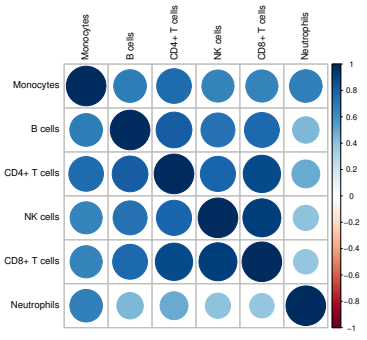



**Fig. 7.** Correlation plot of select GSE28490 samples by cell

## 5 Future Work

It would be interesting to see this comparison done with more algorithms such as CIBERSORT and Bseqsc. Additionally, the full range of SDSC using an ensemble with subject data would certainly increase its robustness.

The combination of datasets to provide more accurate data could be exciting – especially if it was used to expand the number of cells surveyed.

## 6 Figures and Tables

Additional figures and tables are provided below.

## References

Allantaz, F. *et al.* (2012). Expression profiling of human immune cell subsets identifies mirna-mrna regulatory relationships correlated with cell type specific expression. *PloS one*, **7**(1), e29979–e29979. 22276136[pmid].

Becht, E. *et al.* (2016). Cancer immune contexture and immunotherapy. *Current Opinion in Immunology*, **39**, 7–13. Lymphocyte development and activation * Tumour immunology.

Dong, M. *et al.* (2021). Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, **22**(1), 416–427. 31925417[pmid].

DREAM (2019). Tumor deconvolution dream challenge. Available at https://www.synapse.org/#!Synapse:syn15589870/wiki/.

Du, X. *et al.* (2006). Genomic profiles for human peripheral blood t cells, b cells, natural killer cells, monocytes, and polymorphonuclear cells: Comparisons to ischemic stroke, migraine, and tourette syndrome. *Genomics*, **87**(6), 693–703.

Jew, B. *et al.* (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, **11**(1), 1971.

Sung, H. *et al.* (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, **n/a**(n/a).

Wang, X. *et al.* (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, **10**(1), 380.