## Testing Effect of Unintended Bulk RNA-Seq Sample

**Test A: Reproduce exact PDAC Example**

**Test B: Test running PDAC example, but with gene symbols also present in demo1.**

I have uploaded the file that I used for the bulk RNA-Seq sample for **Test B** as *pdac_bulk-MATCHING_SUBSET.csv*. This was created with the uploaded *subset_files.R* script.

**Test C: Test running demo1 bulk RNA-Seq with matched symbols to PDAC example (and all other PDAC files exactly)**

The original goal of this test was to try and assess the amount of dependency on the scRNA-Seq reference. For example, if irrelevant bulk RNA-Seq still looks similar to intended sample, then I might be concerned about using an independent sample of similar type (whose difference I would assume should be between an adjacent sample from same collection and an unrelated bulk RNA-Seq sample).

I have uploaded the file that I used for the bulk RNA-Seq sample for **Test C** as *demo1_bulk-FALSE_PDAC_LABEL-MATCHING_SUBSET.csv*. This was also created with the uploaded *subset_files.R* script.

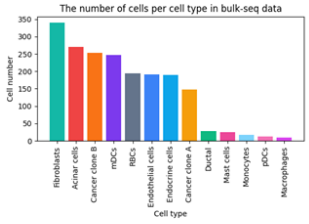
These are all run when commenting out certain variable definitions in the attached file (**run_bulk2space_tests.py**: essentially, the PDAC demo code with varied bulk RNA-Seq input).

If there are any output files that are small enough to share on GitHub, then I would be happy to do so.

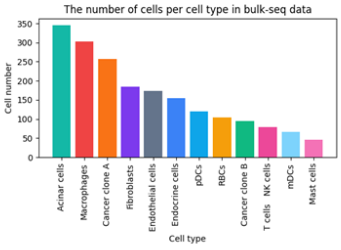
Based upon **Figure 3** in the paper, I think looking at the **cancer cells ("A" and "B" clones)** and the **ductal cells** might be the easiest to qualitatively compare by eye. Or, at least that is what I could see most clearly between the different parts of **Figure 3f**.

When using matching symbols, the "Ductal" assignments in the same PDAC bulk RNA-Seq sample are lost. When providing an irrelevant bulk RNA-Seq sample from "Demo1"), one of the two cancer cell populations remain and there is not a uniform distribution of cell types:

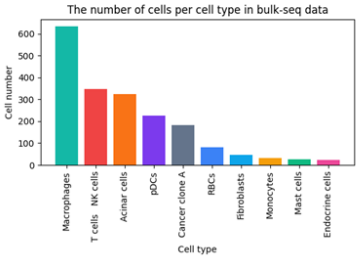
**Reproduce PDAC**
(all full files)




**Demo1 Matching Symbol PDAC**
(full PDAC scRNA-Seq and full PDAC ST)




**Matching Symbol Demo1 Bulk RNA-Seq**
(full PDAC scRNA-Seq and full PDAC ST)

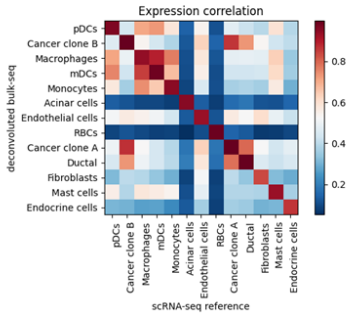



**Lost "Ductal" cell population**

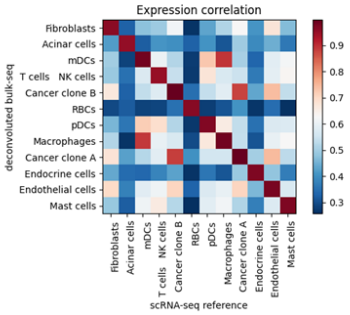
To be fair, there are shifts in the distributions of the cell types for the irrelevant bulk RNA-Seq dataset and an additional loss of the "endothelial cell" population. However, my qualitative impression in the cell type shifts was not that much larger when using an irrelevant bulk RNA-Seq sample than when using a set of matched gene symbols for the same provided PDAC bulk RNA-Seq sample.

I am not sure of all the possible implications, but the higher diagonal correlations still exist when irrelevant bulk RNA-Seq data is provided:

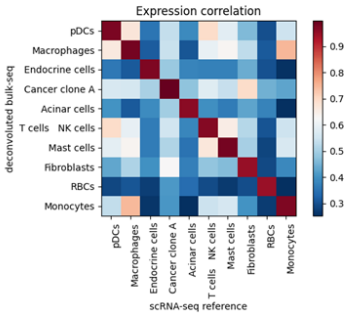
**Reproduce PDAC**
(all full files)




**Demo1 Matching Symbol PDAC**
(full PDAC scRNA-Seq and full PDAC ST)

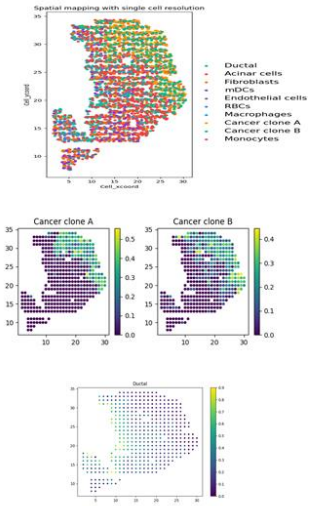



**Matching Symbol Demo1 Bulk RNA-Seq**
(full PDAC scRNA-Seq and full PDAC ST)

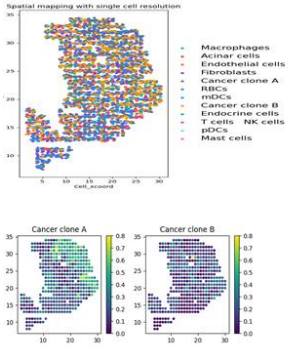



The most clear separation of cells is for the original example. However, there is some preservation of the location of the cancer cells (even arguably when an irrelevant bulk RNA-Seq sample is provided):

**Reproduce PDAC**
(all full files)

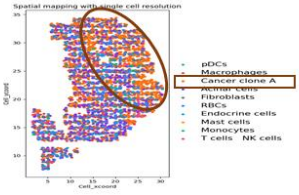



**Demo1 Matching Symbol PDAC**
(full PDAC scRNA-Seq and full PDAC ST)




[lost Ductal cells]

**Matching Symbol Demo1 Bulk RNA-Seq**
(full PDAC scRNA-Seq and full PDAC ST)




[lost *Cancer clone B* cells]

[lost Ductal cells]

While I won't show the images for the results, I also ran analysis with 1000 Epochs instead of 3500 Epochs, and these are the **minimum losses** reported in the log files:

| | **3500 Epochs** (*minimum loss*) | **1000 Epochs** (*minimum loss*) |
|---|---|---|
| **Test A: Exact PDAC Example** | 0.4976 | 0.6726 |
| **Test B: PDAC Match Symbol** (bulk RNA-Seq) | 0.8096 | 1.2149 |
| **Test C: Demo1 Match Symbol** (bulk RNA-Seq) | 0.8758 | 1.2218 |

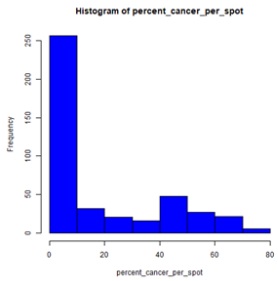
To add some quantification, here are the percentage of cancer cells (for either clone, using *calculate_percent_cancer.R*):

| | **scRNA-Seq** | **ST** |
|---|---|---|
| **Test A: Exact PDAC Example** | 400 / 1,926 (**20.8%**) | 796 / 4,159 (**19.1%**) |
| **Test B: PDAC Match Symbol** (bulk RNA-Seq) | 352 / 1,927 (**18.3%**) | 1,706 / 4,139 (**41.2%**) |
| **Test C: Demo1 Match Symbol** (bulk RNA-Seq) | 184 / 1,927 (**9.5%**) | 1,816 / 4,155 (**43.7%**) |

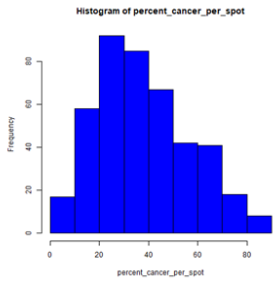
Visually, I did not think it looked like ~40% of spots are mostly cancer cells when using the demo1 bulk RNA-Seq sample. However, the Spatial Transcriptomic (ST) counts have more than one value per spot. For example, if 1-3 out of 10 cells assigned per spot were essentially randomly a cancer cell, then that would make the overall percentage higher than the percentage that you might estimate from ST alone.

Using *calculate_percent_cancer.R*, here is the frequency of cancer cells per spot:

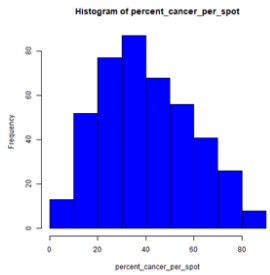
**Reproduce PDAC**
(all full files)




**Demo1 Matching Symbol PDAC**
(full PDAC scRNA-Seq and full PDAC ST)




**Matching Symbol Demo1 Bulk RNA-Seq**
(full PDAC scRNA-Seq and full PDAC ST)




So, there is a shift to increase the frequency of random cancer cells per spot when the matching gene symbols are used (for either the intended PDAC bulk RNA-Seq sample or the Demo1 bulk RNA-Seq sample).

Additionally, I see that **Figure S4** is meant to cover troubleshooting for over-fitting. However, I apologize that I am not sure if I completely understand the logic in showing a lack of over-fitting, even if I might see how random noise could be somewhat like providing an irrelevant bulk RNA-Seq dataset and trying to see if the results look *less* like the scRNA-Seq reference (perhaps, as a best case scenario, providing

bulk RNA-Seq that is not a good fit might generate mixing somewhat like the *uniform* distribution in *Figure S6*?). I am also not sure if this excludes the possibility that over-fitting could occur in some other way that might not be as well captured by that method of adding noise?

For example, I would consider feature selection upstream of cross validation something that would over-estimate the accuracy of a model (and the cross-validation models would not directly be the model to apply in an independent dataset). If all samples were reprocessed from raw data, then I hope having the same set of gene symbols could help avoid the possibility that including only certain gene symbols might impact the estimated accuracy of the method.

**In short, it looks like using a subset of gene symbols has a noticeable effect, but there is some additional divergence when irrelevant bulk RNA-Seq data is provided**. So, I can see some evidence of what I expect should happen with irrelevant data, but I also think there is some structure related to the training dataset (e.g. the scRNA-Seq and Spatial Transcriptomics **reference sample** selection) that impacts the results.