Brain cell type proportion analysis using BRETIGEA

Andrew McKenzie, Minghui Wang, Bin Zhang

2018-09-11

Contents

1 Background for BRETIGEA

Several comprehensive RNA-seq experiments in different brain cell types have now been published in humans and mice. Some of these experiments have profiled gene expression of cell populations isolated through immunopanning procedures. Immunopanning involves immunoprecipitation of particular cell types in cell culture plates, based on selection for an antibody adsorbed to the plate surface. Others studies have performed RNA profiling of single cells with microfluidics devices and used clustering methods to identify cell types from the resulting RNA expression profiles. The devices used for single cell RNA sequencing (scRNA-seq) often select cells based on size or via encapsulation in a droplet and involve the creation of a cDNA library from the transcriptome from a theoretical maximum of one cell.

Existing studies have been mainly based on individual datasets, and are therefore subject to systematic noise, including sampling bias due to sample collection or preparation technique, as well as stochasticity in gene expression. As an increasing number of RNA-seq cell type-specific transcriptomic experiments have become available for both human and mouse, we set out to conduct a comprehensive meta-analysis of brain cell type gene signatures, which is now published in McKenzie et al (2018), doi:10.1038/s41598-018-27293-5. We created cell type-specific (i.e. marker) gene signatures for six cell types: astrocytes (ast), endothelial cells (end), microglia (mic), neurons (neu), oligodendrocytes (oli), and oligodendrocyte precursor cells (opc). The goal of our cell type specifity measure, which is fully described in our manuscript, is to measure whether a gene is expressed in only one cell type relative to the others.

The five data sets used in the creation of the cell type marker signatures can be found in the manuscript.

2 Introduction to BRETIGEA

A major goal of **BRETIGEA** (BRain cEll Type specIfic Gene Expression Analysis) is to simplify the process of defining your own set of brain cell type marker genes by using a well-validated set of cell type-specific marker genes derived from both immunopanning and single cell microfluidic experiments, as described in

McKenzie et al (2018), doi:10.1038/s41598-018-27293-5. There are brain cell type markers available that have been developed from human data, mouse data,, and combinations using data from both species (the default). Notably, if you use your own marker data, the functions in **BRETIGEA** are applicable to bulk gene expression data from any tissue. This vignette shows how you can perform cell type proportion estimation and adjustment on your own bulk gene expression data.

3 Data loading and input format

First, we will load the package and read in example bulk RNA-sequencing data from four brain regions (frontal white matter, temporal cortex, parietal cortex, and hippocampus), which was generated by the Allen Brain Atlas ("Allen Institute for Cell Science. Aging, Dementia and TBI," n.d.) and filtered to contain primarily brain marker genes. We also will load a data frame with additional immunohistochemistry quantification measurements from each brain sample, to use as a validation of the method.

```
library(BRETIGEA, quietly = TRUE)
library(knitr) #only used for vignette creation
```

Here is the format of the inputs:

```
str(aba_marker_expression, list.len = 5)
```

```
## 'data.frame':    395 obs. of  377 variables:
##  $ X488395315: num  0.6557 4.5264 0 0 0.0397 ...
##  $ X496100277: num  0.0951 8.8558 0 0 0.0165 ...
##  $ X496100278: num  0 4.87 0 0 0 ...
##  $ X496100279: num  0 4.85 0 0 0.17 ...
##  $ X496100281: num  0 3.6 0 0 0 ...
##   [list output truncated]
```

```
str(aba_pheno_data, list.len = 5)
```

```
## 'data.frame':    377 obs. of  4 variables:
##  $ structure_acronym.x: chr  "TCx" "FWM" "FWM" "TCx" ...
##  $ ihc_iba1_ffpe      : num  0.0371 0.044 0.0465 0.074 0.1124 ...
##  $ ihc_gfap_ffpe      : num  0.0218 NA 0.0664 0.0181 0.0756 ...
##  $ id                 : chr  "X488395315" "X496100277" "X496100278" "X496100279" ...
```

4 Relative cell type proportion estimation

To run the brain cell type proportion estimation analysis and extract the matrix of surrogate proportion variables for each of the major six brain cell types (astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and OPCs), run this:

```
ct_res = brainCells(aba_marker_expression, nMarker = 50)
kable(head(ct_res))
```

	ast	end	mic	neu	oli	opc
X488395315	-0.0409765	-0.0468875	-0.0249076	0.0226400	-0.0194737	-0.0287028
X496100277	0.0391782	0.0090563	-0.0012271	-0.1361360	0.1323645	0.1322346
X496100278	0.0742051	0.0864415	0.1158266	-0.1360790	0.1534334	0.1555192
X496100279	-0.0091306	-0.0055174	0.0103811	0.0680277	-0.0194953	-0.0216833
X496100281	0.1136897	-0.0070804	0.0825388	0.0116946	-0.0243035	-0.0278465
X496100283	-0.0440731	-0.0263346	-0.0356047	0.0449777	-0.0220543	-0.0188682

4.1 Selecting the nMarker parameter

Note that the above analysis uses $nMarker = 50$ marker genes. A notable trade-off in the selection of the number of marker genes to include in the analysis is that the more marker genes you use, the more likely you are to average out any cell type-specific expression changes that may occur across groups in your sample. On the other hand, the fewer marker genes you use, the higher-quality these marker genes will tend to be in terms of strength of cell type specificity. We have chose $nMarker = 50$ because it has been a reasonable number in our experince, but the goals of your analysis may differ and you may want to choose a different number of marker genes for each cell type.

Note that only marker genes which have been measured in your data set will be used by the cell type proportion estimates, so if your data set has fewer gene measurements (e.g., in a proteomics data set), that may be a reason to use fewer marker genes.

Comparing these cell type proportion estimates to the independent immunohistochemistry quantifications of two marker genes (IBA1 and GFAP), you can see that the correlation is strong.

```
cor_mic = cor.test(ct_res[, "mic"], as.numeric(aba_pheno_data$ihc_iba1_ffpe),
  method = "spearman")
print(cor_mic)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ct_res[, "mic"] and as.numeric(aba_pheno_data$ihc_iba1_ffpe)
## S = 5350800, p-value = 1.729e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.328793
```

```
cor_ast = cor.test(ct_res[, "ast"], as.numeric(aba_pheno_data$ihc_gfap_ffpe),
  method = "spearman")
print(cor_ast)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ct_res[, "ast"] and as.numeric(aba_pheno_data$ihc_gfap_ffpe)
## S = 3591900, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4751708
```

The default cell type proportion estimation method is singular value decomposition, but if you want to use PCA, that is an option as well.

```
ct_res = brainCells(aba_marker_expression, nMarker = 50, species = "combined",
  method = "PCA")
kable(head(ct_res))
```

	ast	end	mic	neu	oli	opc
X488395315	-772.2007	-11.216558	-27.520538	117.72323	-652.7520	-52.75793
X496100277	738.3115	2.166464	-1.355803	-707.87803	4436.8175	243.05687
X496100278	1398.3931	20.678778	127.977252	-707.58137	5143.0376	285.85561
X496100279	-172.0668	-1.319885	11.470150	353.72941	-653.4776	-39.85540

	ast	end	mic	neu	oli	opc
X496100281	2142.4780	-1.693788	91.197438	60.80918	-814.6443	-51.18396
X496100283	-830.5568	-6.299840	-39.339784	233.87439	-739.2522	-34.68110

The *species* argument controls which species the marker genes are derived from, and can be set to "human" and "mouse" for data specific to those species.

If you want to only estimate the proportion of particular cell types, you can do so by setting the *celltypes* argument. Here, we only estimate the proportions of astrocytes, neurons, and oligodendrocytes. Note that the estimates of each cell type is done independently, so choosing to estimate the proportions of one cell type or not will not affect the estimates of the other cell types.

```
ct_res = brainCells(aba_marker_expression, nMarker = 50, species = "combined",
  celltypes = c("ast", "neu", "oli"))
kable(head(ct_res))
```

	ast	neu	oli
X488395315	-0.0409765	0.0226400	-0.0194737
X496100277	0.0391782	-0.1361360	0.1323645
X496100278	0.0742051	-0.1360790	0.1534334
X496100279	-0.0091306	0.0680277	-0.0194953
X496100281	0.1136897	0.0116946	-0.0243035
X496100283	-0.0440731	0.0449777	-0.0220543

5 Using your own cell type marker genes

If you have access to your own marker genes, you can use the *findCells* function instead. This has the same functionality otherwise; *brainCells* is simply a wrapper function for users who want to use the brain cell type marker genes that are provided by **BRETIGEA**. Note the format of the *markers* data frame: you must have one column with the gene symbol, named *markers*, and one column with the corresponding cell type, named *cell*.

```
str(markers_df_brain)
```

```
## 'data.frame':    6000 obs. of  2 variables:
##  $ markers: chr  "AQP4" "ALDH1L1" "BMPR1B" "SLC14A1" ...
##  $ cell   : chr  "ast" "ast" "ast" "ast" ...
```

```
ct_res = findCells(aba_marker_expression, markers = markers_df_brain, nMarker = 50)
kable(head(ct_res))
```

	ast	end	mic	neu	oli	opc
X488395315	-0.0409765	-0.0468875	-0.0249076	0.0226400	-0.0194737	-0.0287028
X496100277	0.0391782	0.0090563	-0.0012271	-0.1361360	0.1323645	0.1322346
X496100278	0.0742051	0.0864415	0.1158266	-0.1360790	0.1534334	0.1555192
X496100279	-0.0091306	-0.0055174	0.0103811	0.0680277	-0.0194953	-0.0216833
X496100281	0.1136897	-0.0070804	0.0825388	0.0116946	-0.0243035	-0.0278465
X496100283	-0.0440731	-0.0263346	-0.0356047	0.0449777	-0.0220543	-0.0188682

6 Adjusting bulk gene expression data for estimated cell type proportions

BRETIGEA also offers users the ability to adjust their original gene expression matrices for the estimated cell type proportion estimates, in order to deconvolute the signal.

```
brain_cells_adjusted = adjustBrainCells(aba_marker_expression,
    nMarker = 50, species = "combined")
expression_data_adj = brain_cells_adjusted$expression
```

Note that *adjustBrainCells* is a wrapper function to *adjustCells* and if you have your own markers (e.g., for a non-brain data set), then you can use that interface instead for deconvolution of more general cell types.

As you can see, following adjustment, there is no longer a correlation between the RNA expression of the microglia marker gene AIF1 and its encoded protein IHC quantification (IBA1), nor between the RNA and protein expression of the astrocyte marker gene GFAP. (Note there *is* a non-significant trend towards a residual correlation here, which may be because GFAP is not a perfect marker of astrocyte proportion in this data set, but instead varies across samples based on disease state, region, and other factors).

```
cor.test(as.numeric(aba_marker_expression["AIF1", ]),
  as.numeric(aba_pheno_data$ihc_iba1_ffpe), method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(aba_marker_expression["AIF1", ]) and as.numeric(aba_pheno_data$ihc_iba1_ffpe)
## S = 5566800, p-value = 5.348e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3017048
```

```
cor.test(expression_data_adj["AIF1", ], as.numeric(aba_pheno_data$ihc_iba1_ffpe),
  method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  expression_data_adj["AIF1", ] and as.numeric(aba_pheno_data$ihc_iba1_ffpe)
## S = 7975600, p-value = 0.9931
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##           rho
## -0.0004520843
```

```
cor.test(as.numeric(aba_marker_expression["GFAP", ]), as.numeric(aba_pheno_data$ihc_gfap_ffpe),
  method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(aba_marker_expression["GFAP", ]) and as.numeric(aba_pheno_data$ihc_gfap_ffpe)
## S = 3582800, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
```

```
## 0.476499
```

```
cor.test(expression_data_adj["GFAP", ], as.numeric(aba_pheno_data$ihc_gfap_ffpe),
  method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  expression_data_adj["GFAP", ] and as.numeric(aba_pheno_data$ihc_gfap_ffpe)
## S = 6458000, p-value = 0.2962
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.05637708
```

7 Help and other resources

If you have any problems with or questions about using this package, please open an issue on Github or contact the package maintainer.

References

"Allen Institute for Cell Science. Aging, Dementia and TBI." n.d. http://aging.brain-map.org/.