Creating Test Data

Kathi Reinisch

The goal is to create small test data sets that are as close to the ground truth from Hoek et al. as possible.

Loading data

You might need to change the path in this chunk of code in the Rmd as I won't include the datasets. Make sure you made the methods from the immunedeconv2 package available somehow!

```
DATADIR = "F:/Katharina/Kathi"
x <- fread(file.path(DATADIR, "maynard_2020_annotated_fine_2k/X_tpm.csv"),
    header = F)
var <- fread(file.path(DATADIR, "maynard_2020_annotated_fine_2k/var.csv"),
    header = T)
obs <- fread(file.path(DATADIR, "maynard_2020_annotated_fine_2k/obs.csv"),
    header = T)
load("F:/Katharina/Kathi/HoekPBMC_gtruth.RData")
load("F:/Katharina/Kathi/Hoek_sample_annotations.RData")

sc <- t(x)
rownames(sc) <- var$symbol
colnames(sc) <- obs$Run

transform_refData <- t(RefData)

source("immunedeconv2/R/deconvolution_algorithms.R")
source("immunedeconv2/R/MOMF.R")
source("immunedeconv2/R/bisque.R")
source("immunedeconv2/R/data_processing.R")
```

Preparing the data

I only included certain cell types, as I thought many cells of one type would be better than only one cell for each cell type. Also, the ground truth only included a few cell types. The cell types included in the test data set are "T cell CD4", "T cell CD8", "T cell dividing", "T cell regulatory", "B cell", "Monocyte conventional", "Monocyte non-conventional", "Macrophage", "NK cell". For the smaller dataset, only "T cell CD4", "T cell CD8", "B cell", "Monocyte conventional", "NK cell" are included.

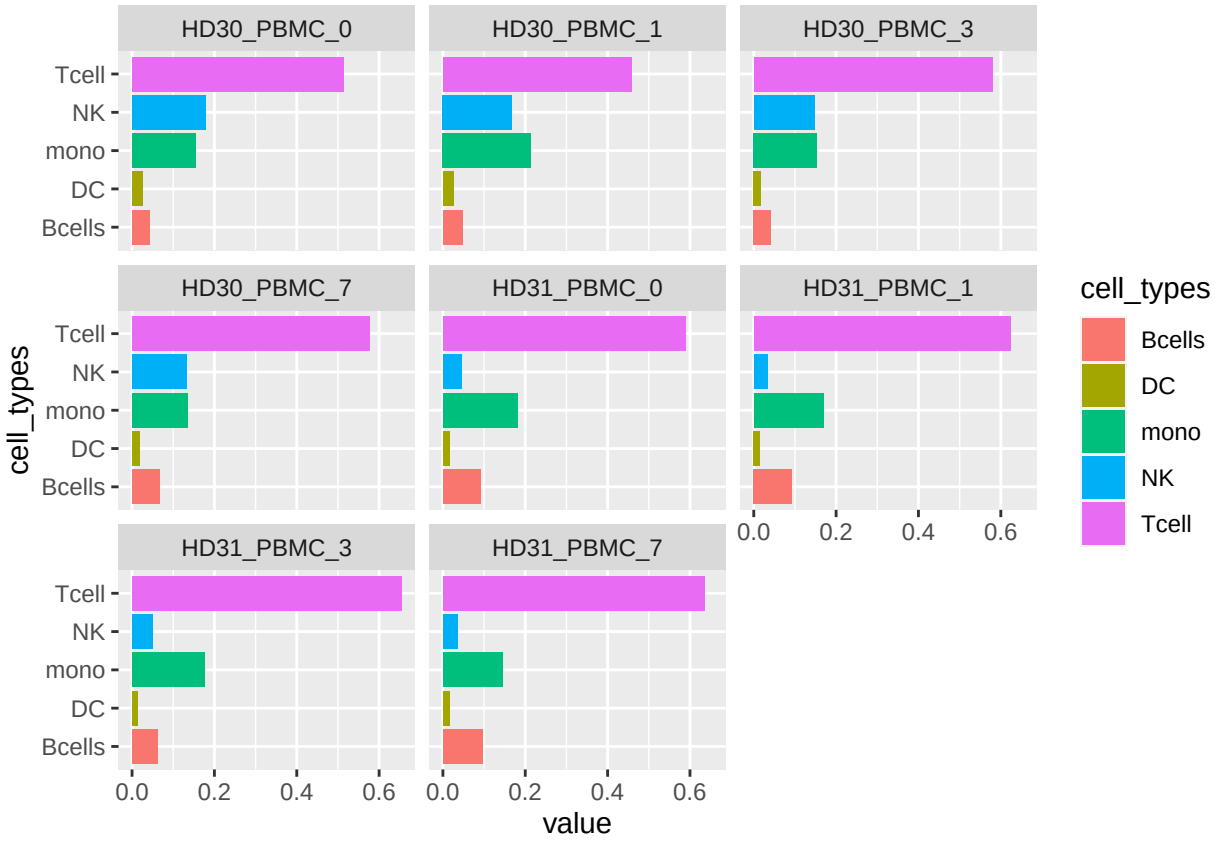
```
contains_rem <- which(obs$cell_type %in% c("T cell CD4", "T cell CD8",
    "T cell dividing", "T cell regulatory", "B cell", "Monocyte conventional",
    "Monocyte non-conventional", "Macrophage", "NK cell"))
obs_rem <- obs[contains_rem, ]
```

```
sc_rem <- sc[, contains_rem]

contains_rem_small <- which(obs$cell_type %in% c("T cell CD4",
    "T cell CD8", "B cell", "Monocyte conventional", "NK cell"))
obs_rem_small <- obs[contains_rem_small, ]
sc_rem_small <- sc[, contains_rem_small]
```

Choosing the best subset

This is kind of what the ground truth looks like:



I decided to make two subsets. One contained all genes that had a positive value (sum of all cell types) in a signature matrix based on the 2k sc data by Maynard et al., which was built by Bisque, intersecting with the bulk data from Hoek et al. This dataset contains 300 cells.

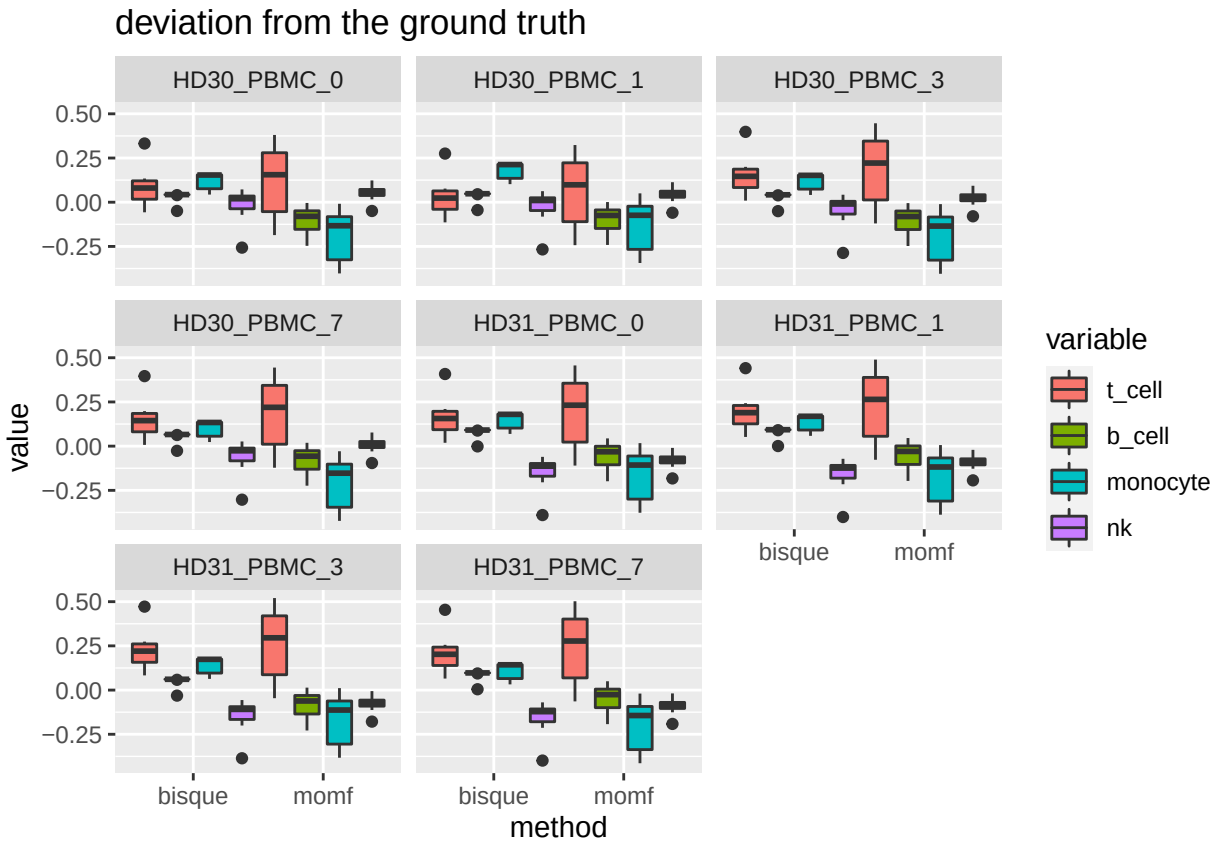
The smaller subset contains only 50 cells and 100 random genes of those fulfilling the conditions described above.

The process to select the cells was the following: I ran deconvolution by MOMF and Bisque on random subsets of 300 cells and calculated a score for the deconvolution results. This score is calculated by the sum of abs(proportion_refdata(celltype)-proportion_deconv(celltype)) for all cell types. The subset with the lowest score was the chosen test data set. Please note: As sampling is random, you might receive different "best data sets" each time you run this! To save the data sets, remove # before the save commands. All these calculations might seem a little crazy, but I needed to make sure i catch every possible error as the process is random and knitting just stops when an error occurs...

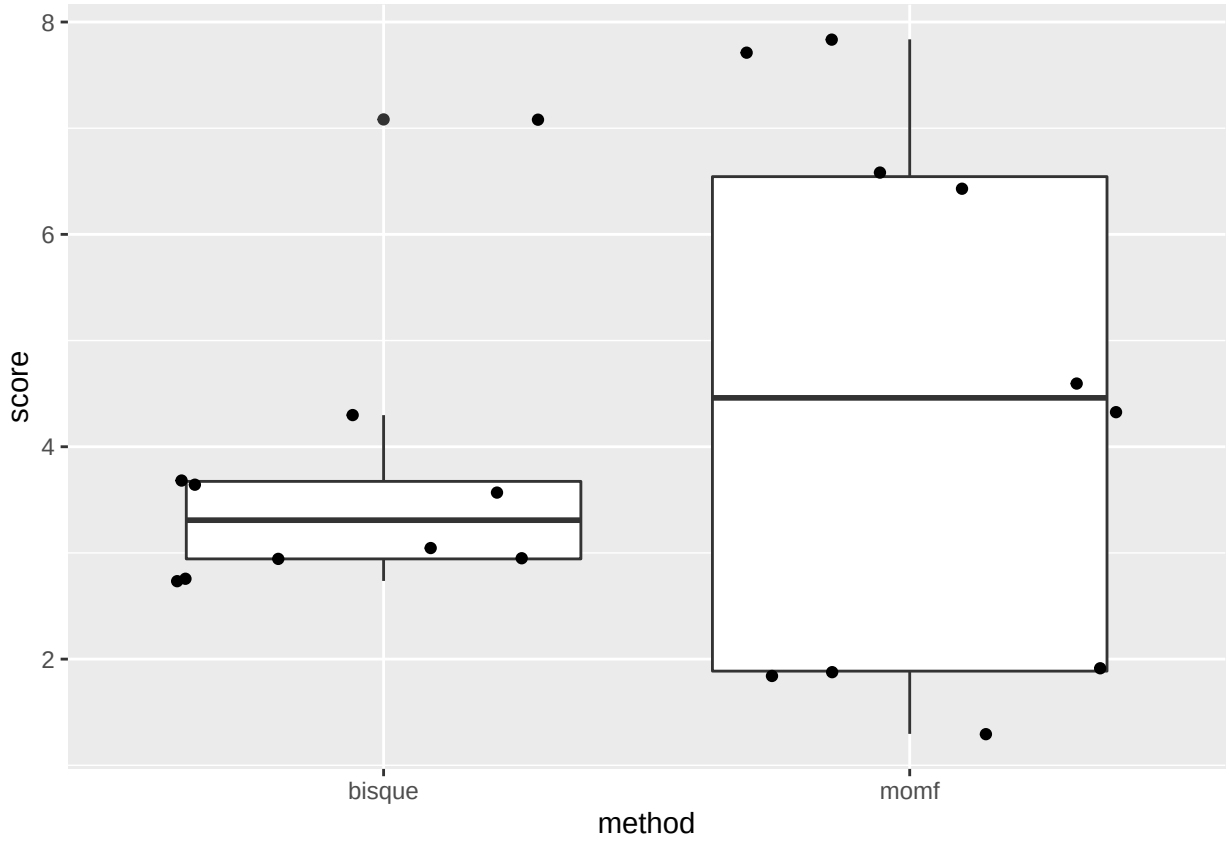
This process may take a while, as MOMF is super slow (about 3mins per deconvolution with about 300 cells)

Below you can see some statistics of how good the subsets did in terms of deconvolution. Remember, the lower the score the closer the deconvolution result is to the ground truth.

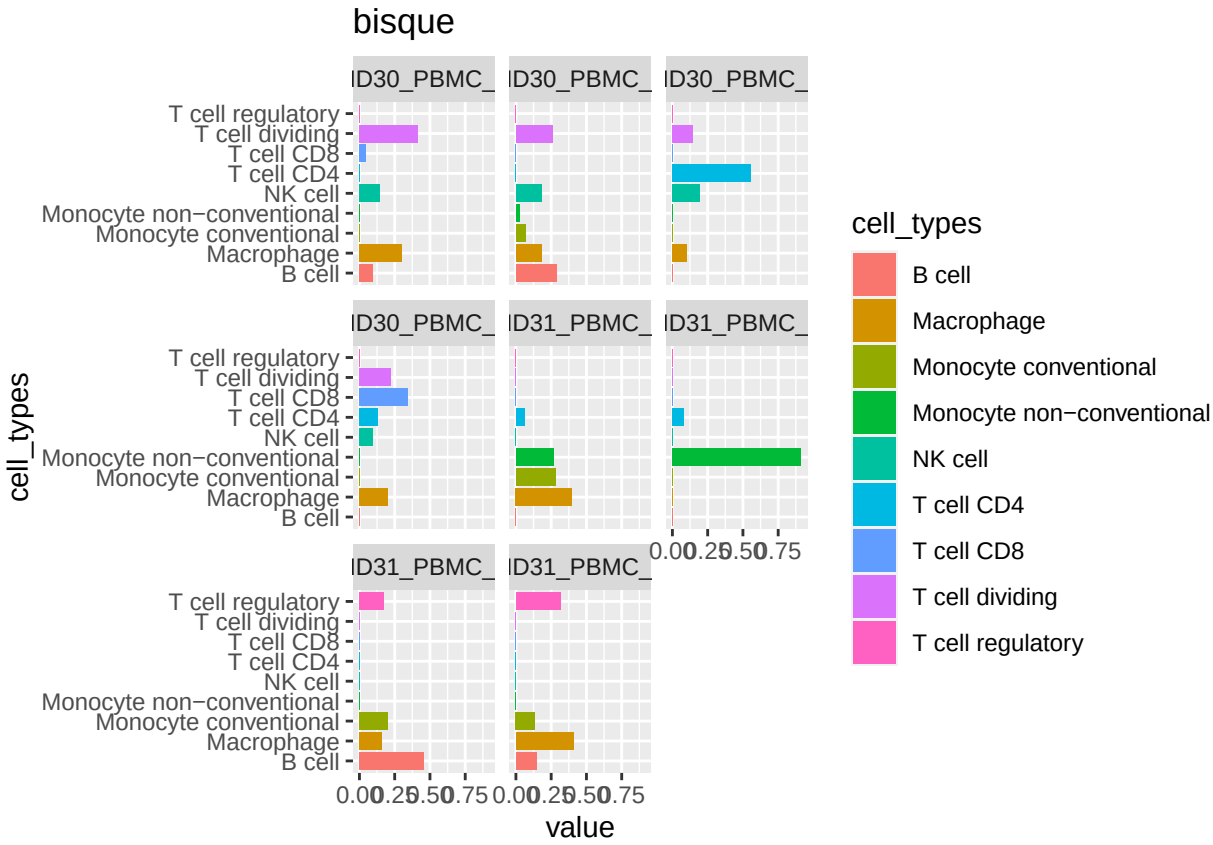
```
melted_li <- melt(li[[1]])
ggplot(melted_li, aes(x = method, y = value, fill = variable)) +
    geom_boxplot(position = position_dodge(1)) + facet_wrap(~sample) +
    labs(title = "deviation from the ground truth")
```

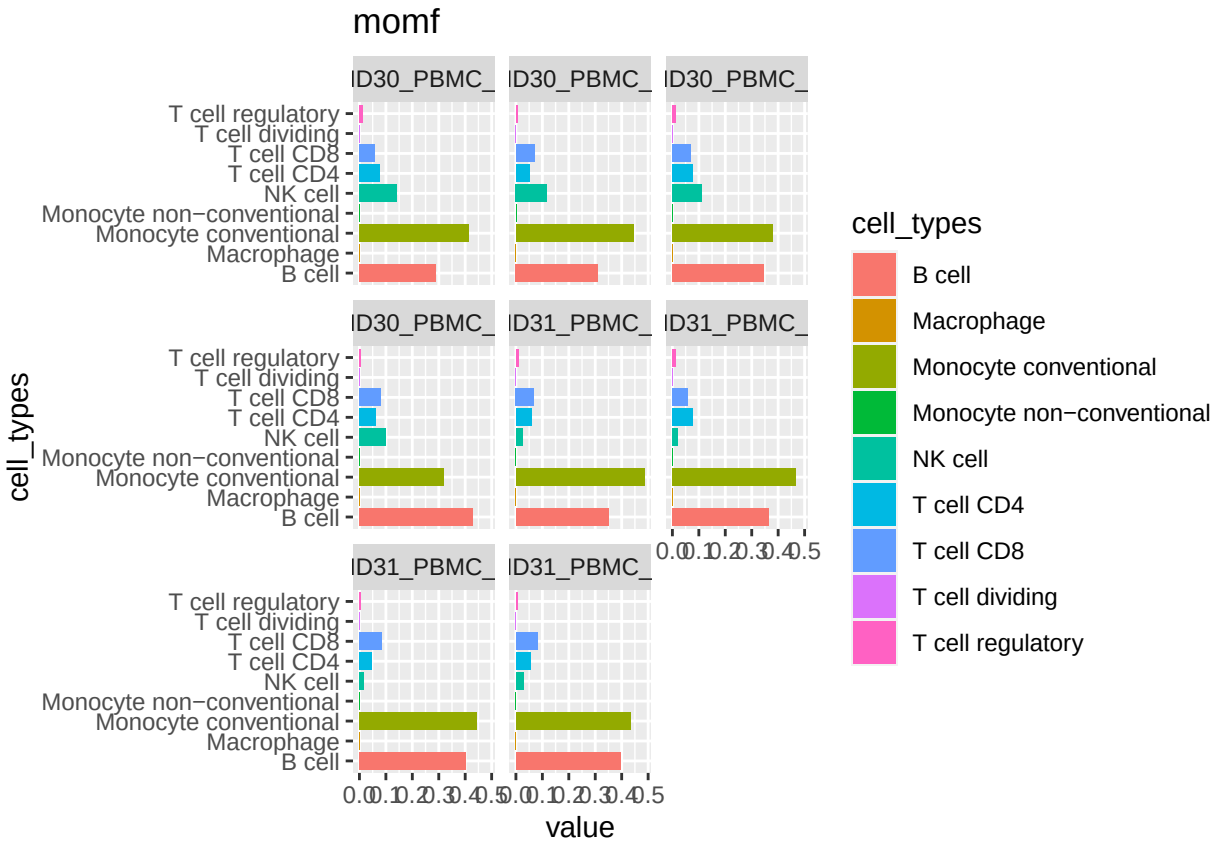



```
ggplot(melt(li[[2]]), aes(x = variable, y = value)) + geom_boxplot() +
    geom_jitter() + labs(y = "score", x = "method")
```



This is how deconvolution with the "best" subset looks for Bisque and MOMF.


bisque
ID30_PBMC_
ID30_PBMC_
ID30_PBMC_
ID30_PBMC_
ID31_PBMC_
ID31_PBMC_
ID31_PBMC_
ID31_PBMC_
T cell regulatory
T cell dividing
T cell CD8
T cell CD4
NK cell
Monocyte non−conventional
Monocyte conventional
Macrophage
B cell
0.00
0.25
0.50
0.75
cell_types
value
cell_types
B cell
Macrophage
Monocyte conventional
Monocyte non−conventional
NK cell
T cell CD4
T cell CD8
T cell dividing
T cell regulatory


momf
ID30_PBMC_
ID30_PBMC_
ID30_PBMC_
ID30_PBMC_
ID31_PBMC_
ID31_PBMC_
ID31_PBMC_
ID31_PBMC_
T cell regulatory
T cell dividing
T cell CD8
T cell CD4
NK cell
Monocyte non−conventional
Monocyte conventional
Macrophage
B cell
cell_types
0.0 0.1 0.2 0.3 0.4 0.5
value
cell_types
B cell
Macrophage
Monocyte conventional
Monocyte non−conventional
NK cell
T cell CD4
T cell CD8
T cell dividing
T cell regulatory