MDI: weeks 1 to 3

Stephen Coleman

20 February 2019

Prelude

Many of the R snippets below require these R packages (Kolde 2019, Dowle and Srinivasan (2019), Neuwirth (2014)):

```
library(pheatmap)
library(data.table)
library(RColorBrewer)
```

MDI

Reading

Inital work involved reading Mason et al. (2016) and reading the manual of the associated program. Expecting that this would have difficulties scaling (based on Paul's beliefs from November) to nine datasets I also looked into Bardenet, Doucet, and Holmes (2017) for general ideas, but more specifically read a good chunk of Betancourt (2017). These topics were of interest as I know the MDI implementation uses a Gibbs sampler - I hoped we could make gains in the computational trtactability of using all 9 datasets in one go with a more clever sampling method.

Data handling

MDI uses data that has common row names across all datasets. The data from the CEDAR cohort is in the form of [people $\times$ probes]. We want to transpose this; to accomplish this task we wrote an Rscript that can be called form the Linux command line with appropriate arguments to transpose .csv files in a given directory. The `data.table` package by Dowle and Srinivasan (2019) was of great use in making this quick to call - for our 9 datasets of ~300 people and 8-16 thousand probes it took less than a minute to convert the files.

Another script was required to remove all the non-ubiquitous probe IDs. Any probe not present in all datasets was removed to allow MDI to perform as MDI cannot handle missing data. This reduced the set of probes to 4,964.

We performed MDI on these 9 reduced datasets with 10,000 iterations, a burn-in of 1,000 and a thinning rate of 25. MDI overfits the clustering problem beginning with 50 clusters; this reduced down to around 10 occupied clusters per dataset. We assumed that the median cluster allocation was the predicted cluster forgetting about the **label-flipping** problem of unsupervised methods.

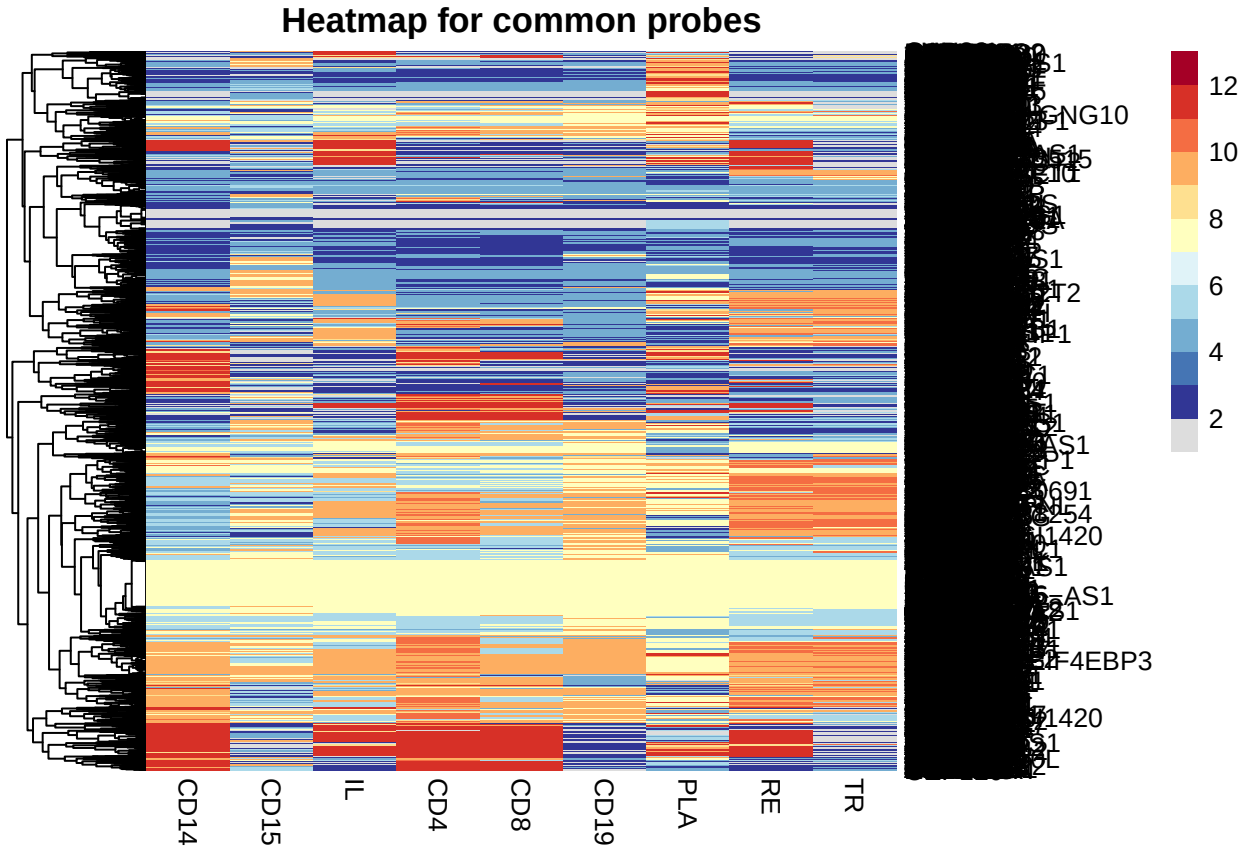
We constructed a dataframe of the median allocation per dataset for each probe. The probes are translated to the associated gene using the key available online. As some of the probes mapped to non-unique genes we use a matrix to hold our data (as this does not require unique row names). We read this in and visually inspect the head of the data.

```
# Read in the data and convert to a matrix with apporpriate row names
compare_df <- data.table::fread("~/Desktop/MDI/Runs/Run 1/allocation_data_run_1.csv")
compare_df_mat <- as.matrix(compare_df[, -10])
row.names(compare_df_mat) <- compare_df$V1
```

```
# Inspect the data
head(compare_df_mat)
```

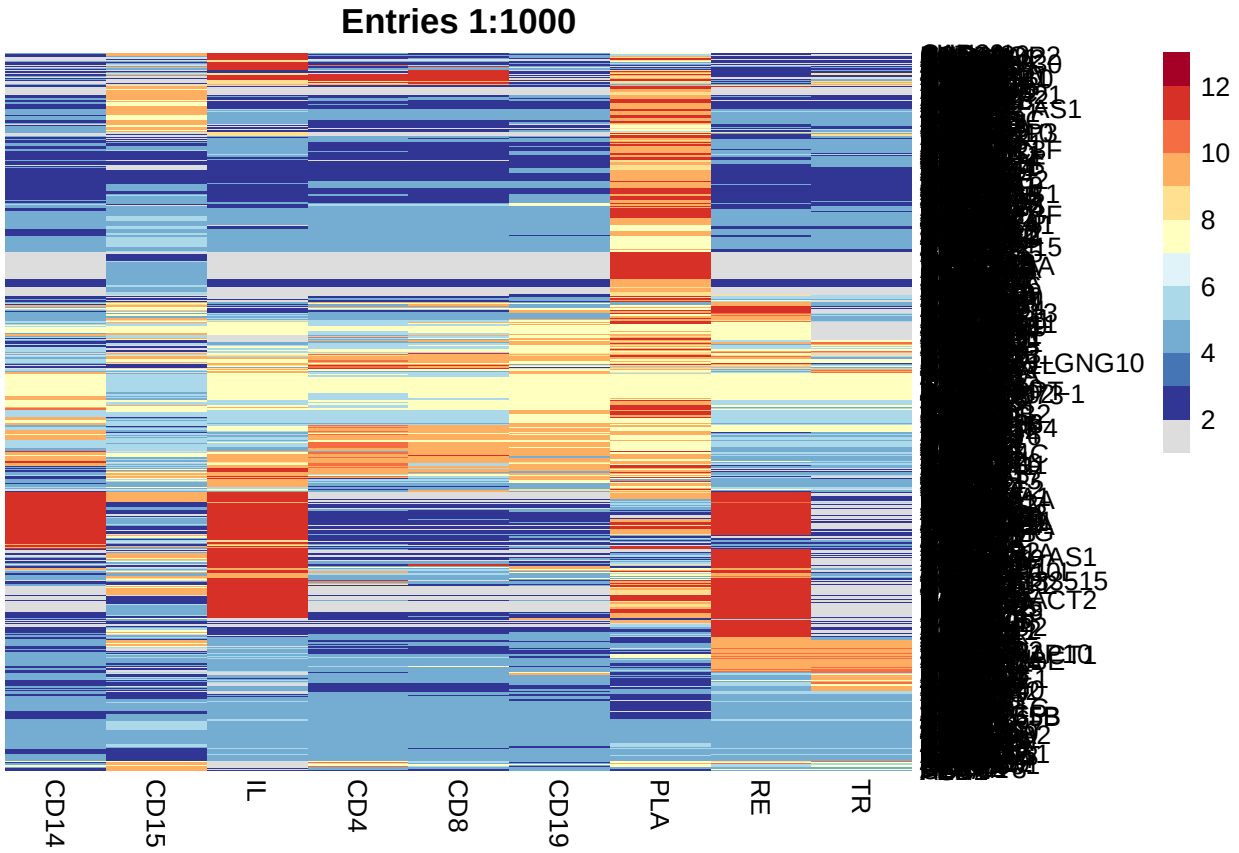
```
##         CD14 CD15 IL CD4 CD8 CD19 PLA RE TR
## GGACT      8    8  8   8   8    8   8  8  8
## A4GNT      5    6 10   6   6   10   5  5  5
## AAAS       3    8  3   3   3    3   5  3  3
## AACS       8   10  8   8   8    8   8  8  8
## AACSP1     8    6  6   6   8    8   5  6  6
## NCEH1     10    8 10  10  10   10   8 11 10
```

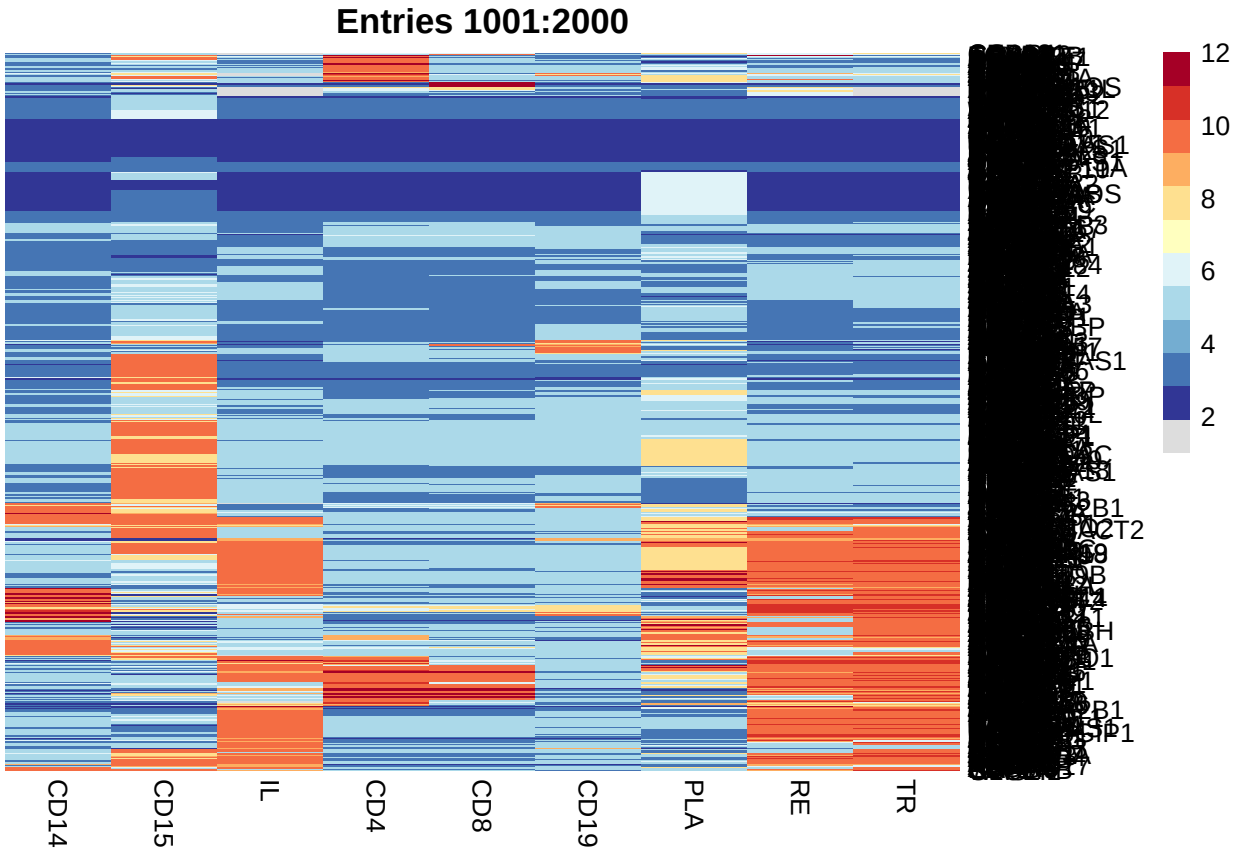
Using this data we generated a heatmap.

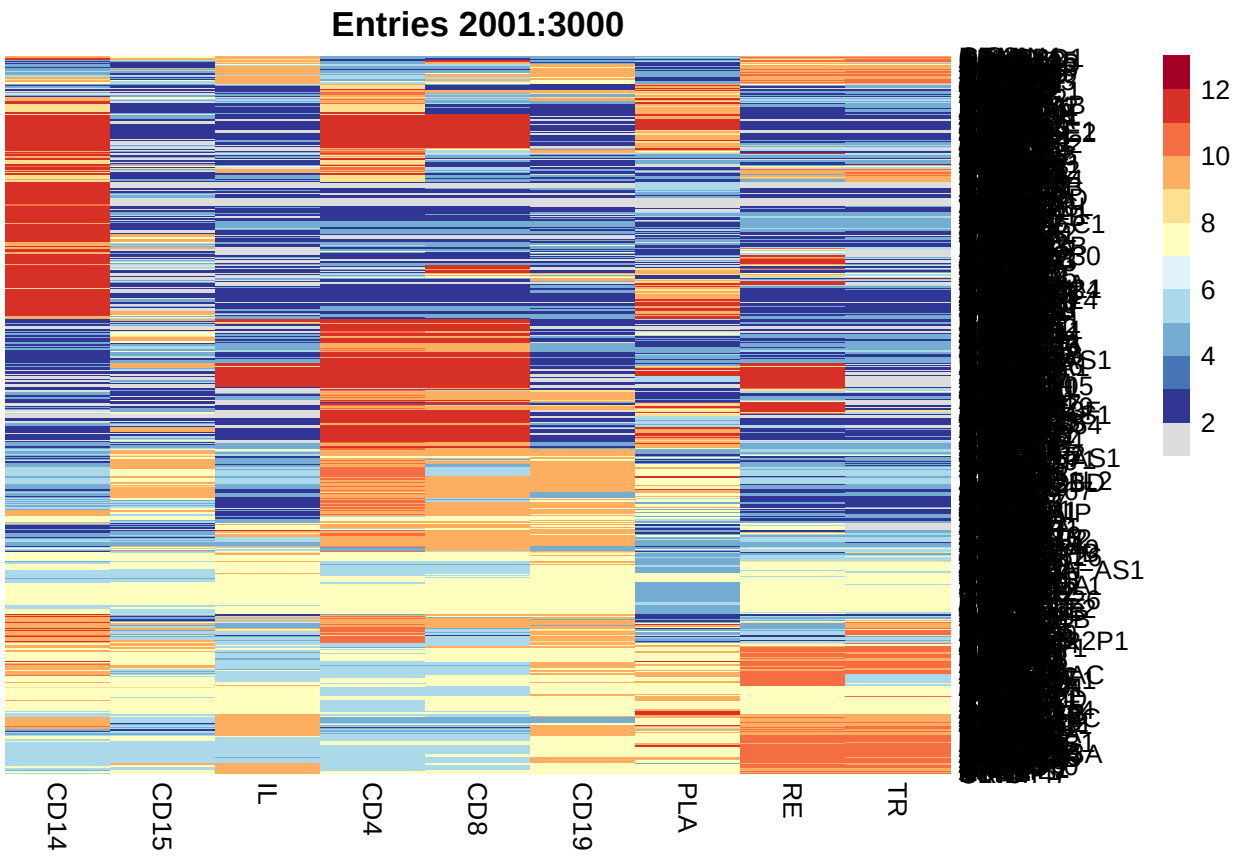


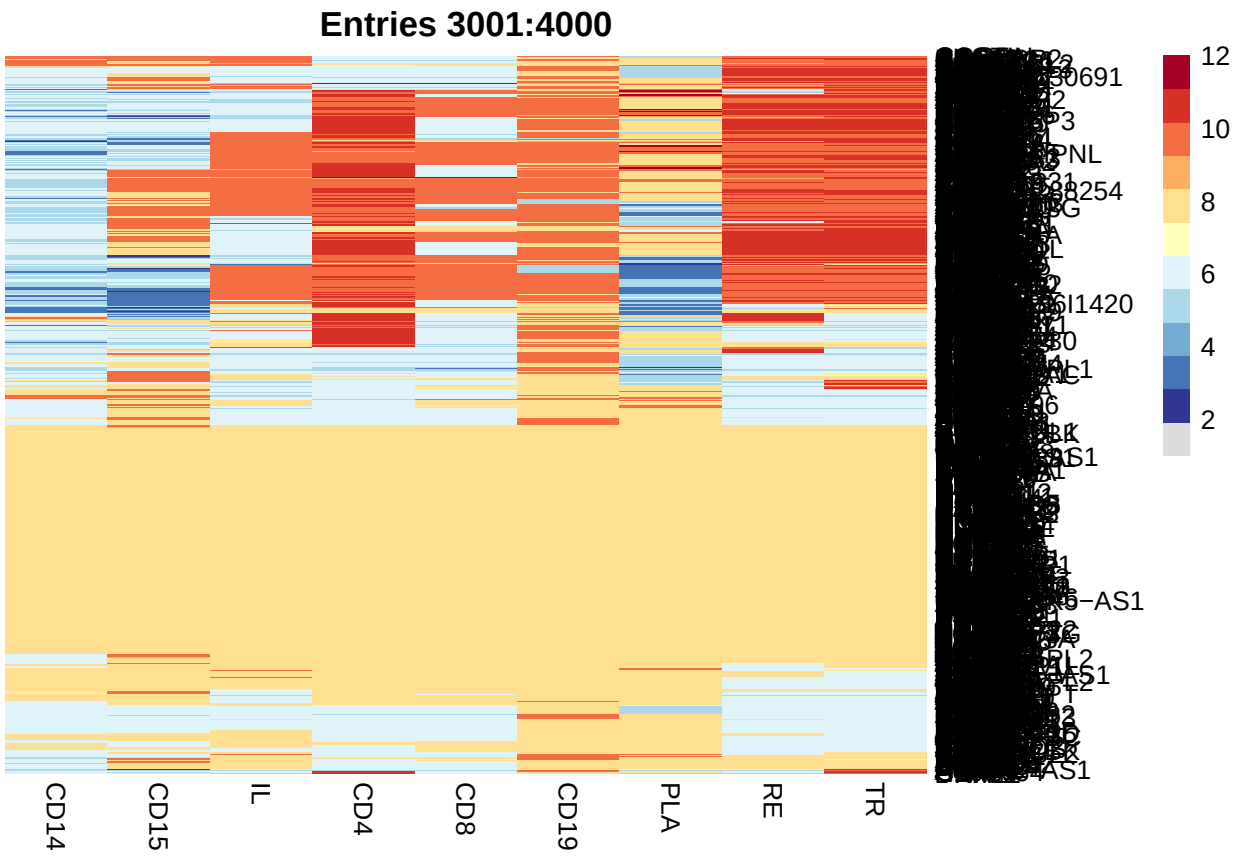
There are 14 clusters present here; to enable comparison we went for a wider colour palette. This is the reason for the garishness of this heatplot. There is appears to have lower similarity than one might expect, particularly considering the selection of data - we only include the probes that are sufficiently expressed in each cell type to register for measurement. Possibly this is an artifact of not considering label-flipping.

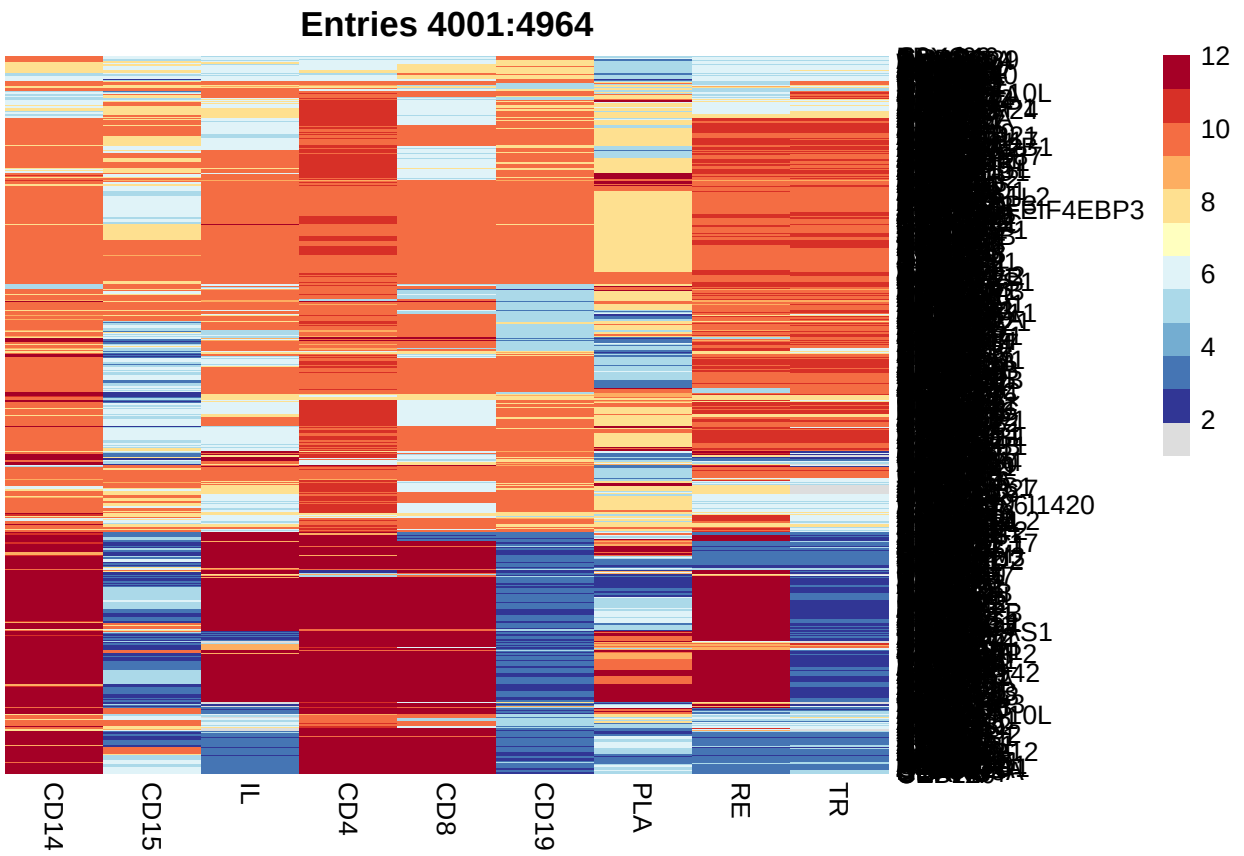
With the same allocation data and maintaining the ordering from the heatmap of the full dataset we inspect subsamples of the data.


Entries 1:1000
CD14
CD15
IL
CD4
CD8
CD19
PLA
RE
TR
12
10
8
6
4
2

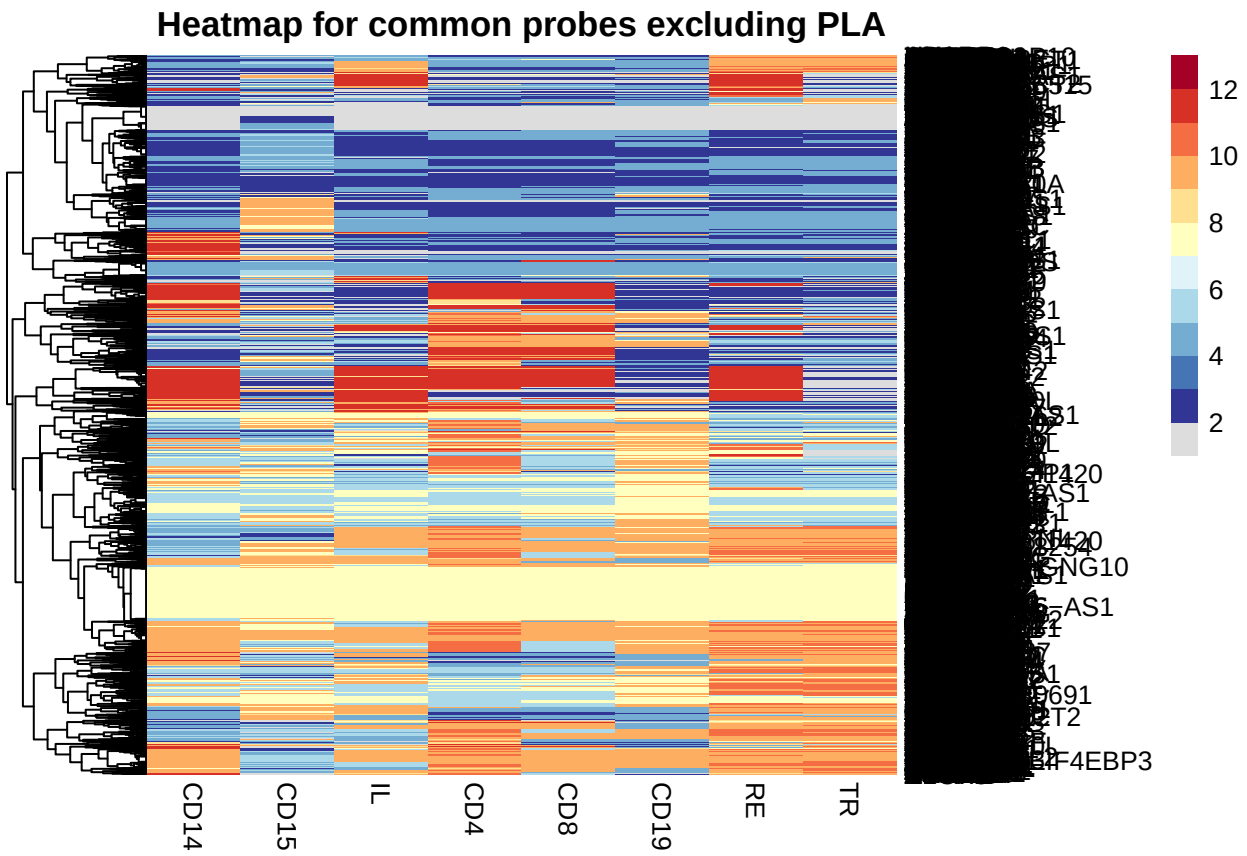

Entries 1001:2000
12
10
8
6
4
2
CD14
CD15
IL
CD4
CD8
CD19
PLA
RE
TR


Entries 2001:3000
CD14
CD15
IL
CD4
CD8
CD19
PLA
RE
TR
12
10
8
6
4
2

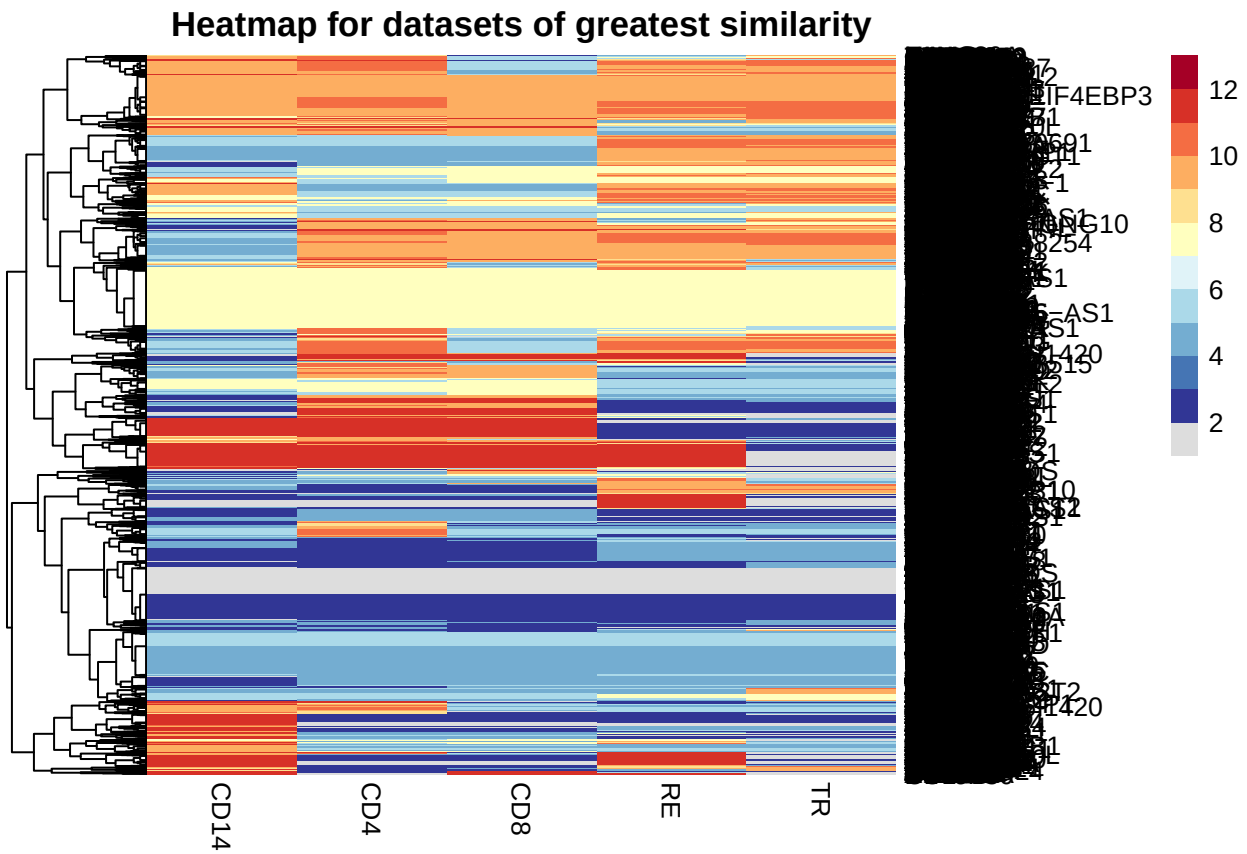

Entries 3001:4000
12
10
8
6
4
2
CD14
CD15
IL
CD4
CD8
CD19
PLA
RE
TR



With this increased granularity, we see that after the first 1,000 rows the similarity across tissues becomes greater. There is still large sections of dissimilarity, but the clustering more coherent than the first plot suggests. Notice also that the PLA dataset (platelets) is the greatest source of dissimialrity - this is also the thinnest dataset with only 8 thousand probes present in the full format.



We consider only four most similar datasets, CD14, CD4, CD8, RE and TR.



Including all probes

References

Bardenet, Rémi, Arnaud Doucet, and Chris Holmes. 2017. "On Markov Chain Monte Carlo Methods for Tall Data." *The Journal of Machine Learning Research* 18 (1). JMLR. org: 1515–57.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

Dowle, Matt, and Arun Srinivasan. 2019. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Kolde, Raivo. 2019. *Pheatmap: Pretty Heatmaps*. https://CRAN.R-project.org/package=pheatmap.

Mason, Samuel A, Faiz Sayyid, Paul DW Kirk, Colin Starr, and David L Wild. 2016. "MDI-Gpu: Accelerating Integrative Modelling for Genomic-Scale Data Using Gp-Gpu Computing." *Statistical Applications in Genetics and Molecular Biology* 15 (1). De Gruyter: 83–86.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.