Bio-IT task with *meth-atlas*

# Introduction

*Meth-atlas* is a bioinformatic tool that performs deconvolution analysis on epigenomic data, specifically on methylome data. It takes a *csv* file with samples data to be analyzed and a reference *csv* file, and returns a new *csv* file with the deconvoluted samples and a stacked bar chart.

Rows in both input files correspond to CpG sites, and their values in each column to the level of DNA methylation in that site (for each sample or for each cell type, depending on the file). The output consists of the proportion of each cell type in each provided sample.

# Objectives

To perform deconvolution analysis on a provided file containing bulk DNA methylation beta values for a series of samples, using a reference that consists of seven blood cell types methylation profiles, using *meth-atlas*. To determine possible sample clusters and compute differential methylation analysis between them.

# Methods

*Meth-atlas* was installed by cloning the corresponding github repository. In order to run it, a conda environment containing all necessary *Python3* packages was installed using a provided *yaml* file. The program was runned via command line, providing the necessary files as arguments.

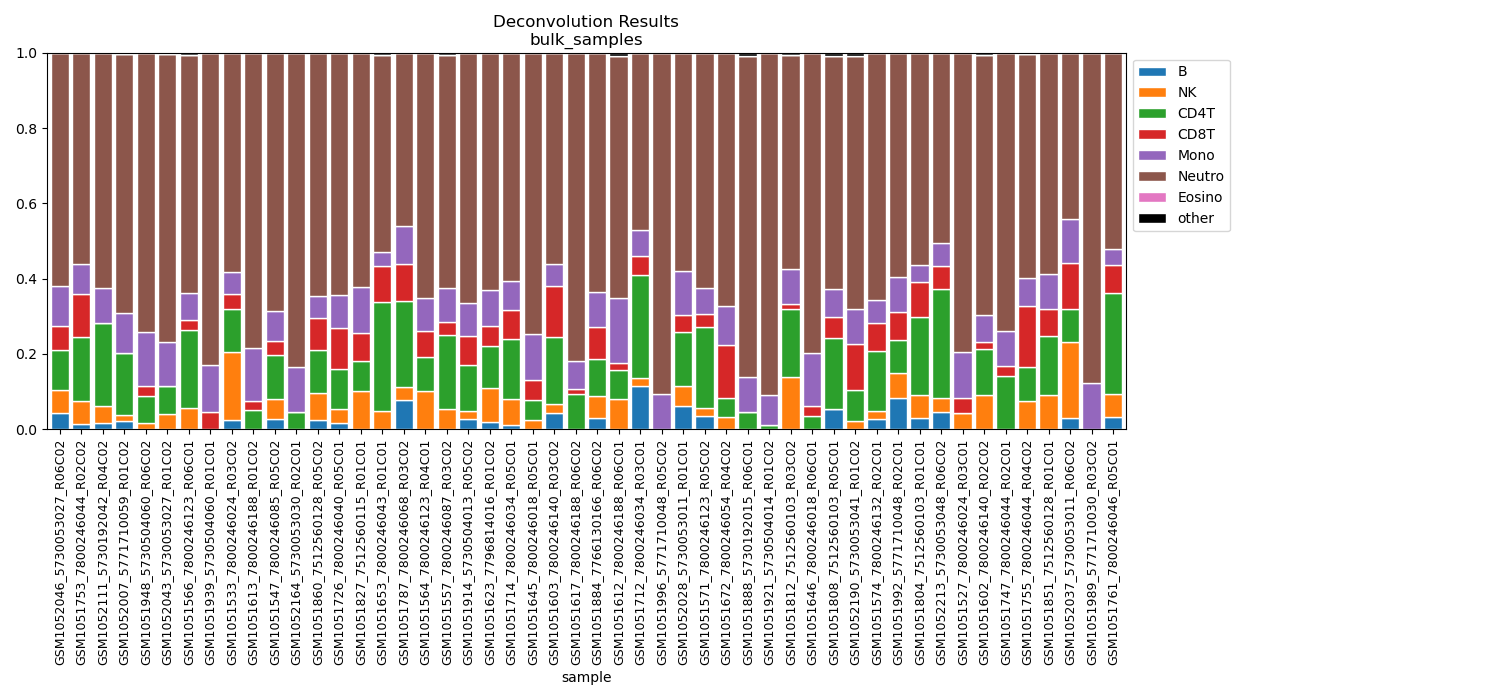
In order to determine possible clusters in the analyzed samples and CpGs, Pearson correlation analysis and PCA were carried out using *scikit-learn*. Next, the potential clusters were used to search for differentially methylated CpGs (DMGs) using *R-limma*. Furthermore, the array’s annotation package in R was used to determine genes associated with those CpGs.

Finally, correlation analysis between the DMGs and the previously determined CpG clusters was done; and enrichment analysis for the DMGs was also carried out to better understand the biology behind.

All scripts, outputs and data are available in the github repository: <https://github.com/CDSchuster/Bio-IT-VIB>.

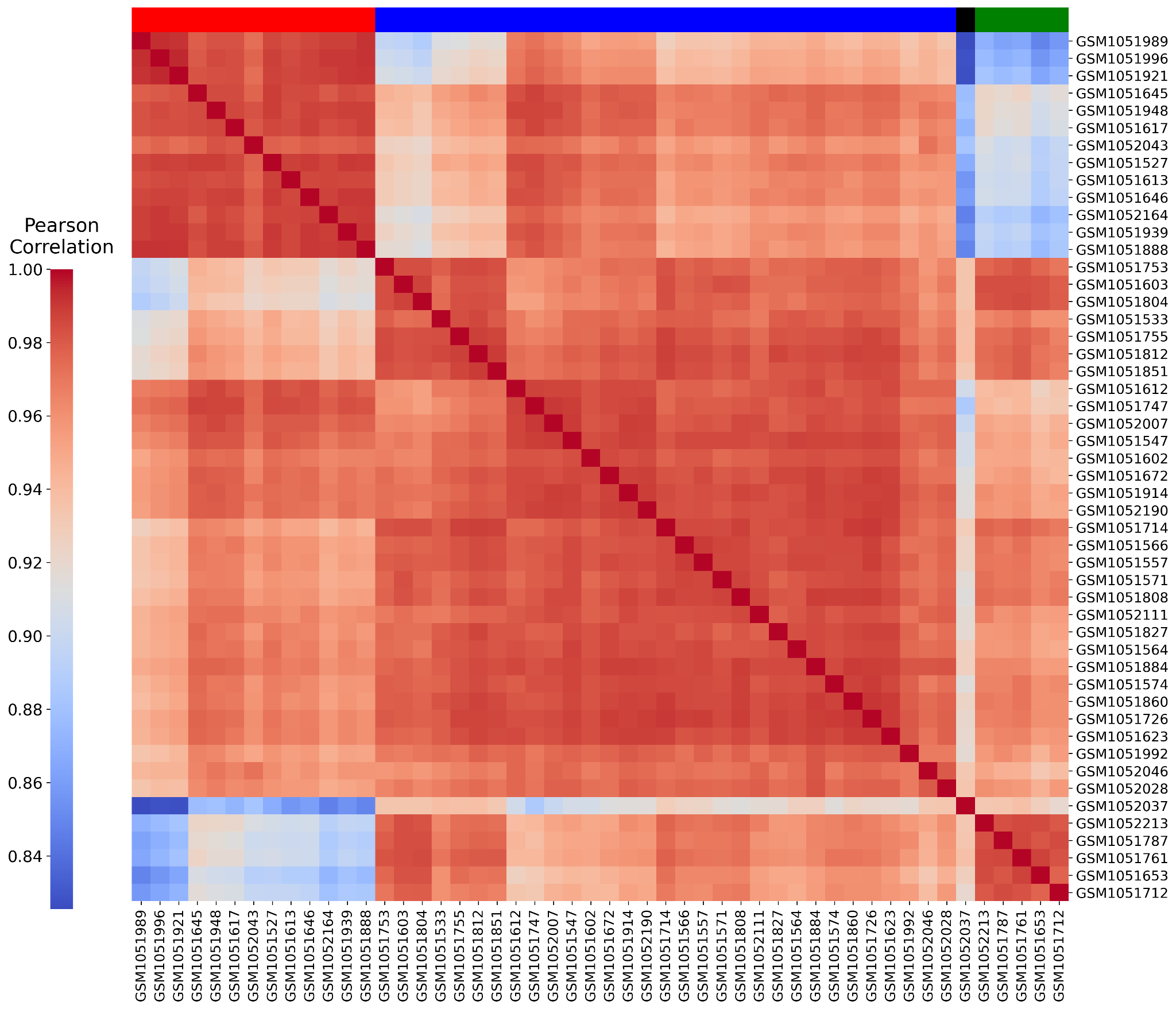
# Results

After running *Meth-atlas*, two files were obtained. One is a *csv* file containing the proportions of each blood cell type for the 50 samples that were provided to analyze. The second file is a stacked bar chart representing the same results that are in the *csv* file, allowing for an easier and more visual way to discuss the data.

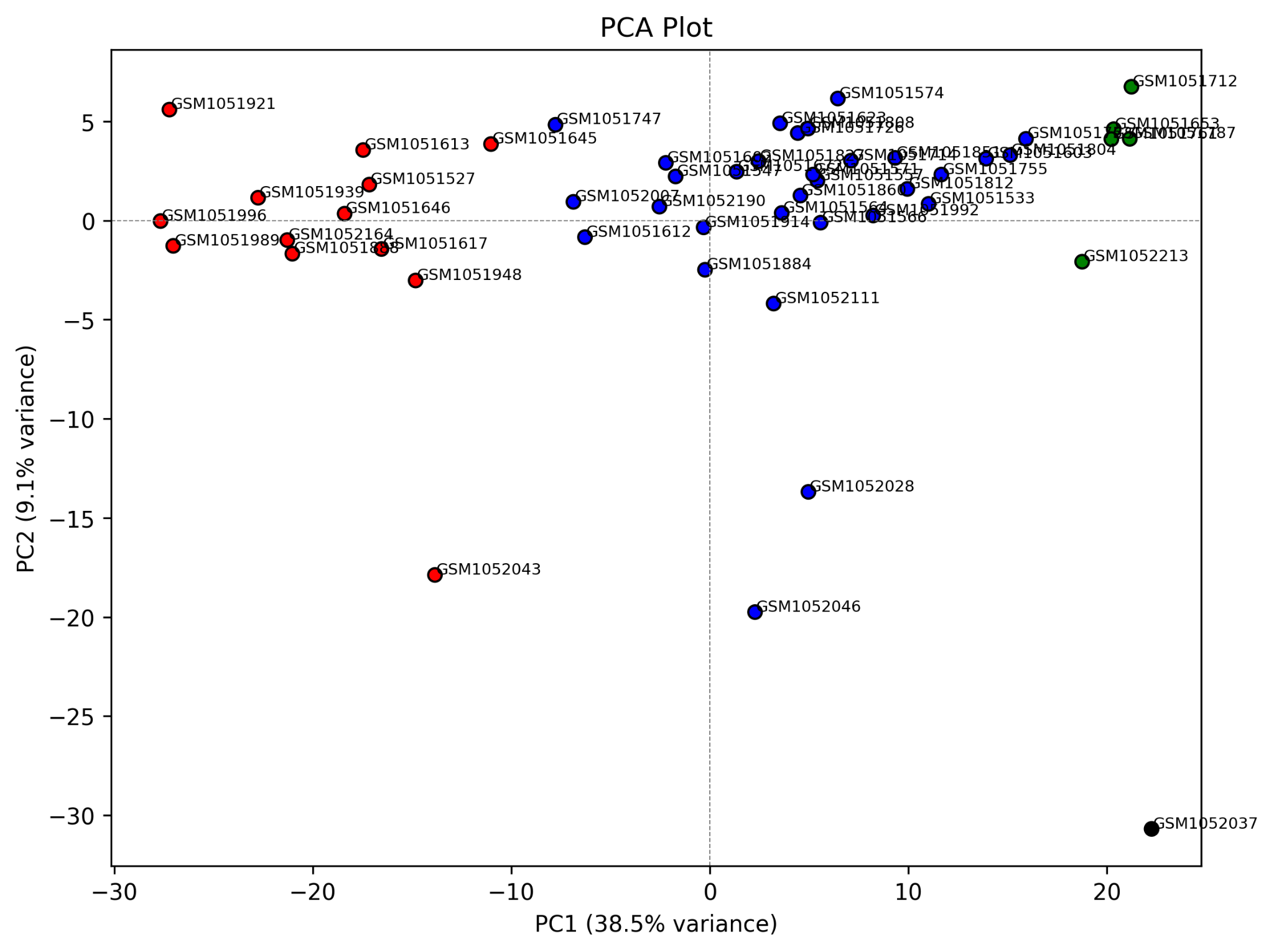


**Fig 1.** Deconvolution results from Meth-atlas for the given DNA methylation data

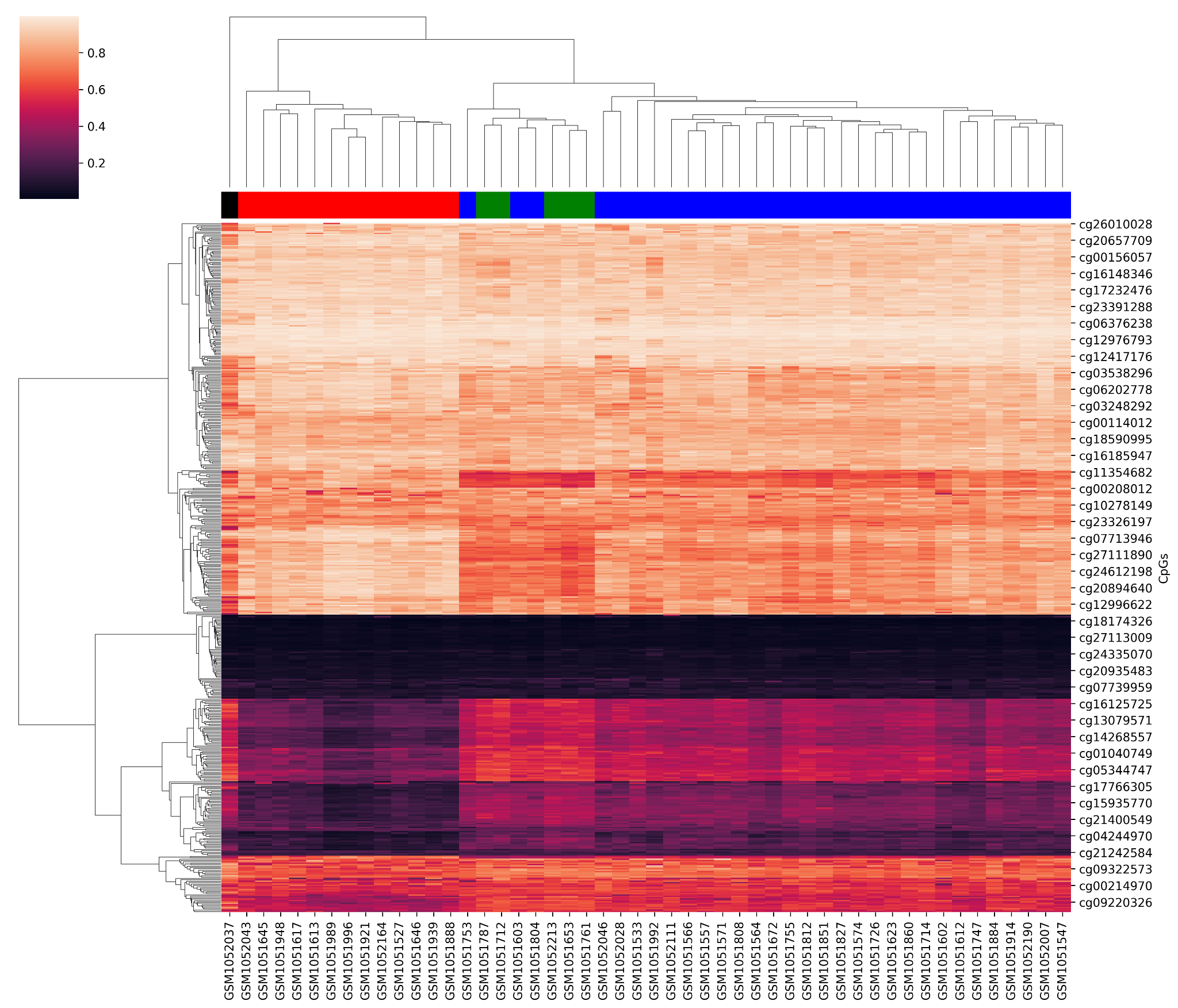
On the other hand, Pearson correlation analysis between the samples (Fig. 2) allowed to determine possible clusters, as it can be seen in the PCA (Fig. 3). These clusters also seem to correlate pretty well with the methylation values (Fig. 4).



**Fig 2.** Pearson correlation results between samples. The color bar on top indicates the clusters that were determined (red, blue and green are clusters 1, 2 and 3, and the black samples is a possible outlier).

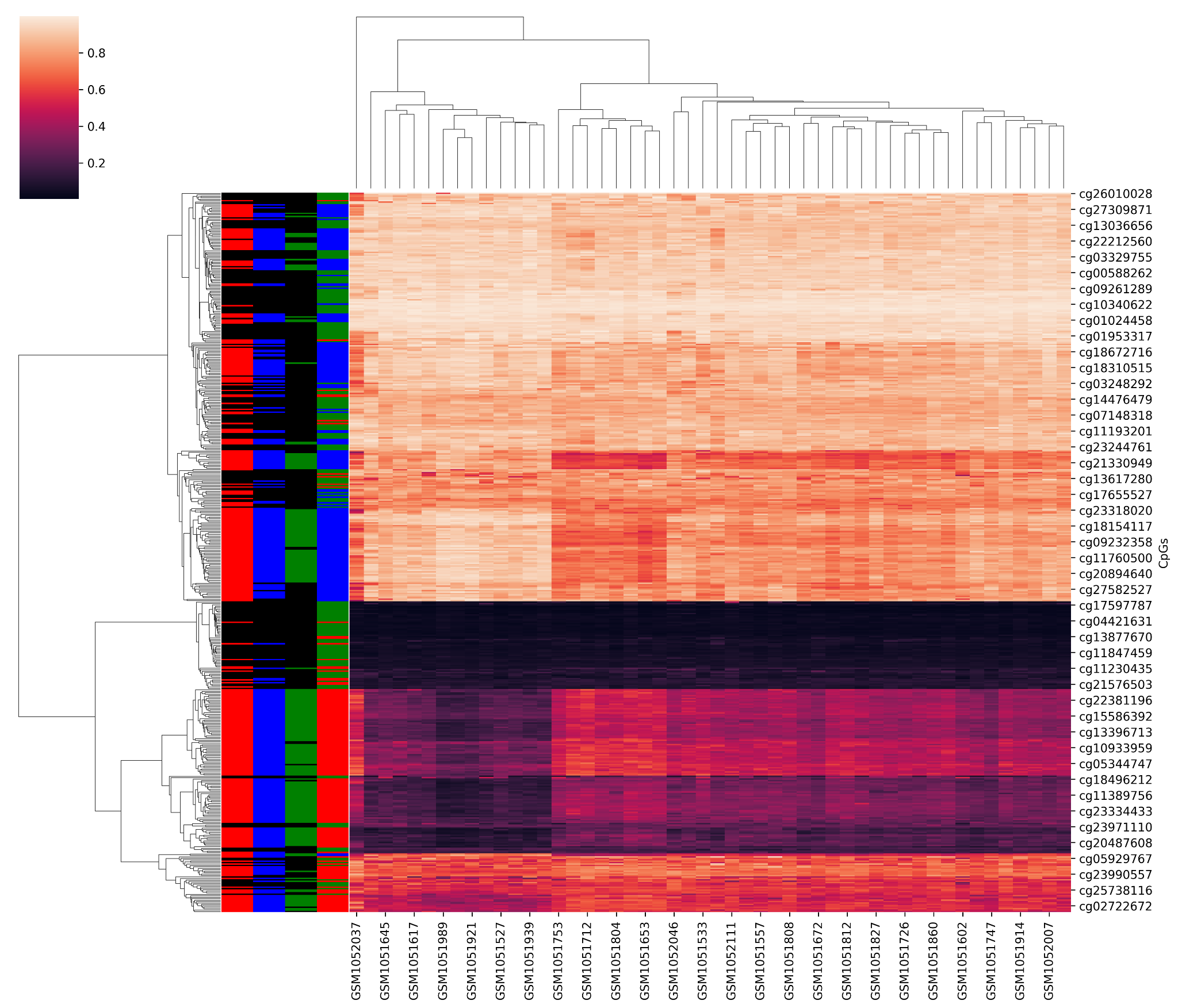


**Fig 3.** PCA for the analyzed samples. The color scheme to indicate the samples’ clusters is the same as in Fig. 2.

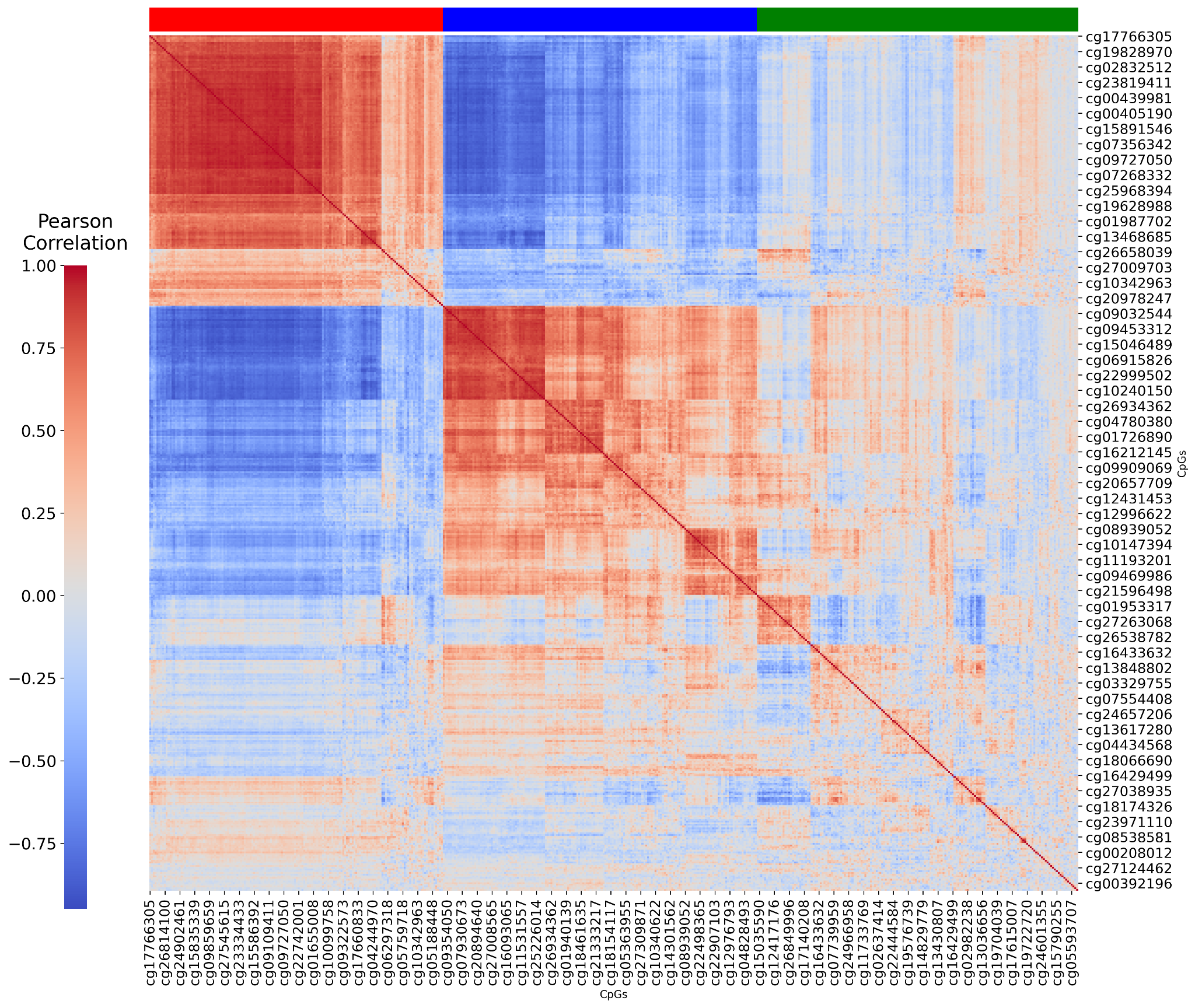


**Fig 4.** Heatmap representing the correlation between the determined clusters and the methylation levels for each sample

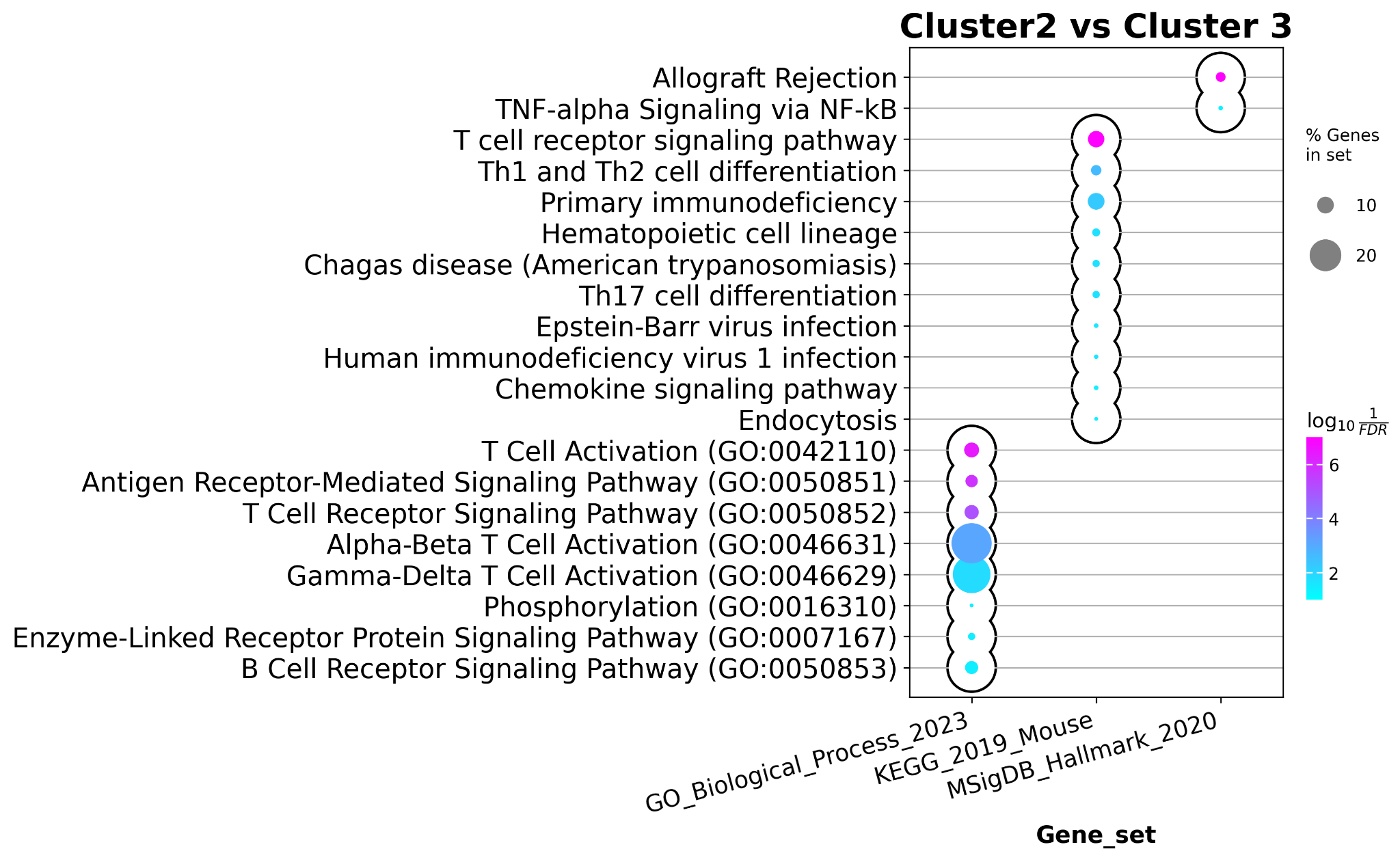
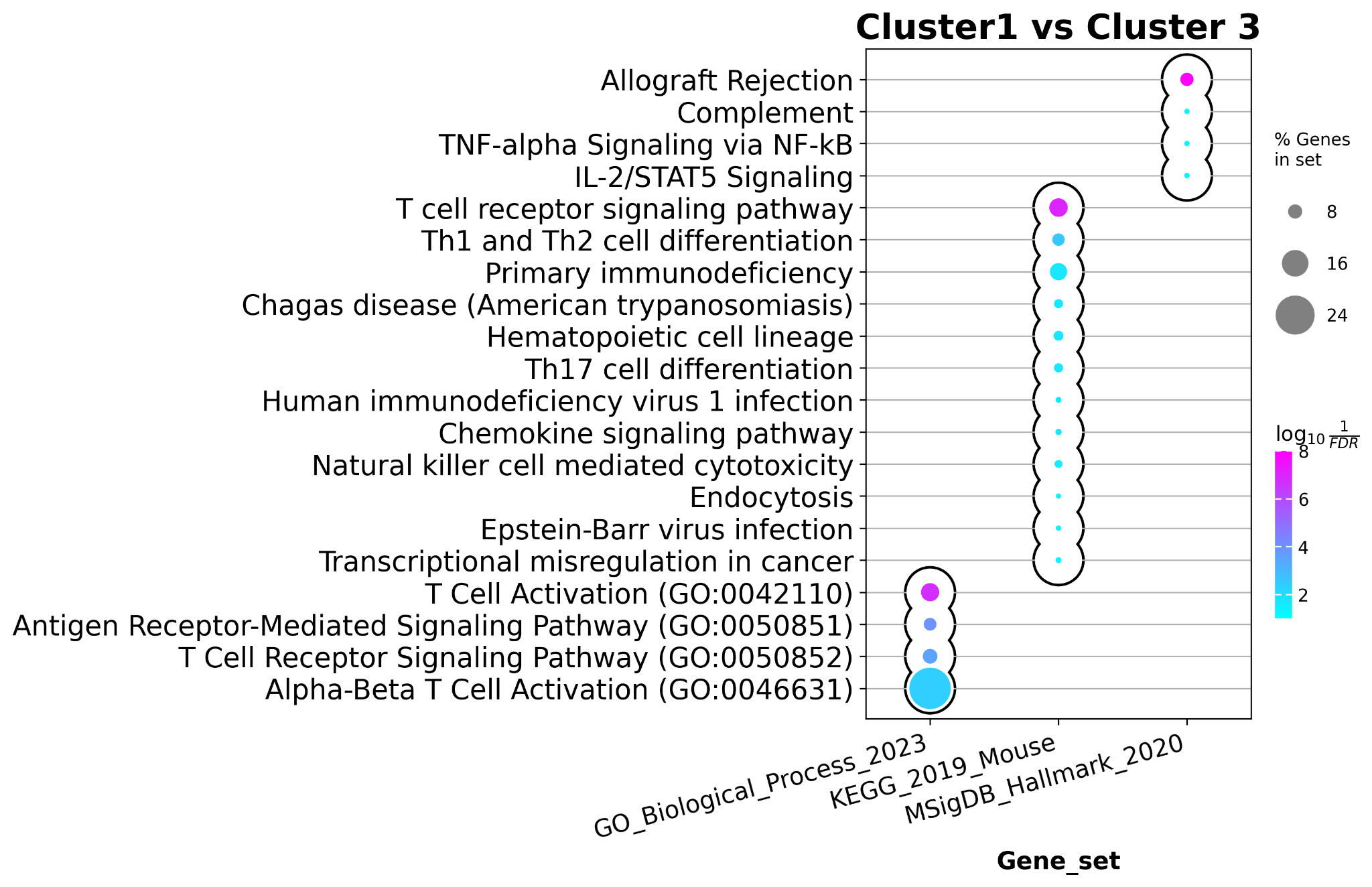
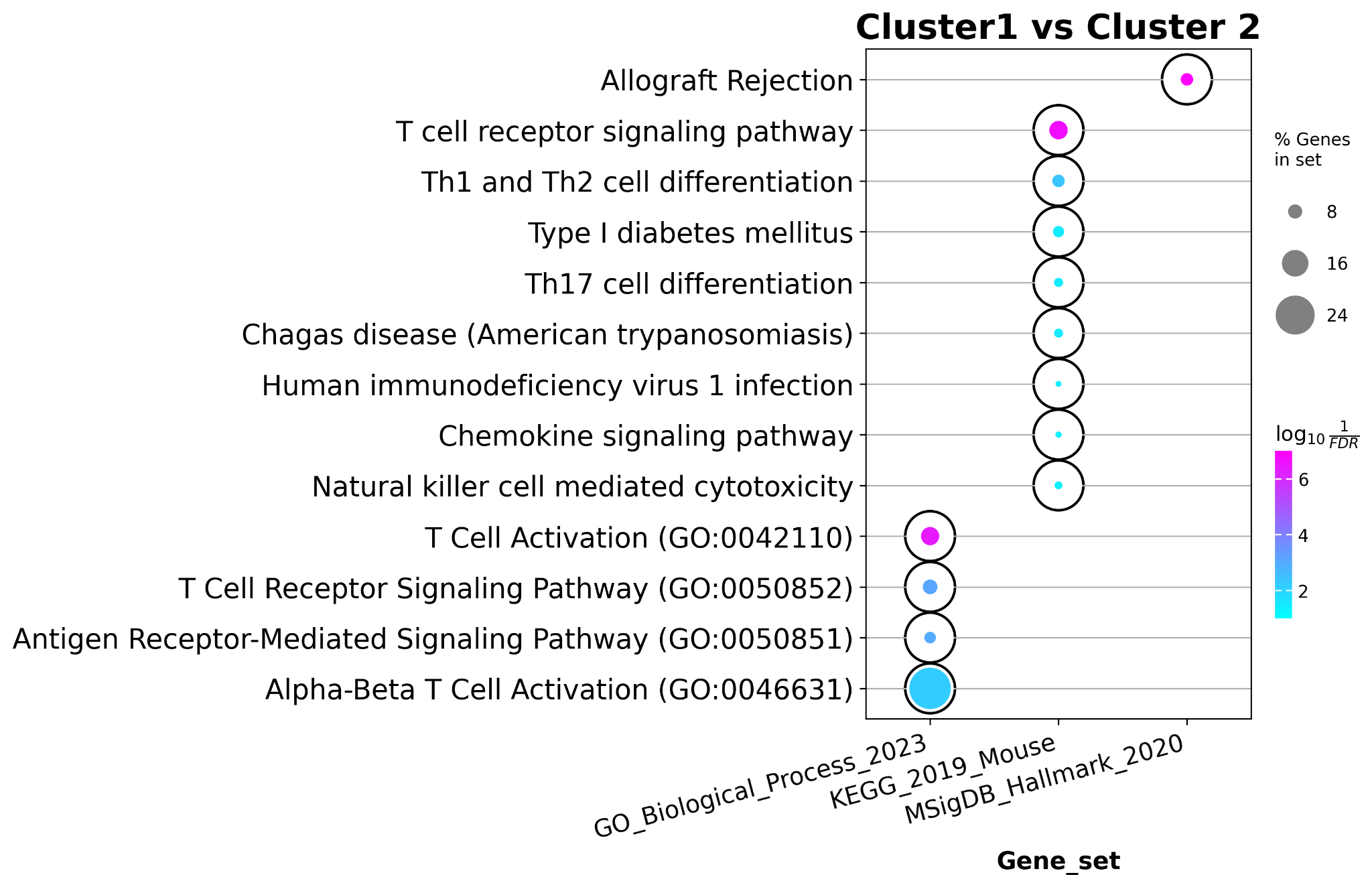
After sample clusters determination, differential methylation analysis was carried out between each pair of clusters (Fig. 5). DMGs were correlated with CpG clusters that had previously determined with Pearson correlation analysis (Fig. 6). Finally, enrichment analysis was performed on each set of DMGs (Fig. 7).



**Fig 5.** Correlation between each set of DMGs, the CpG clusters and the methylation levels. First 3 vertical color bars correspond to the DMG sets cluster1 vs. cluster2, cluster1 vs. cluster3 and cluster2 vs. cluster3 respectively (black indicates a CpG that was not significantly methylated between the two compared clusters). The fourth color bar corresponds to the CpG clusters inferred with Pearson correlation.



**Fig 6.** Pearson correlation between the CpGs, as well as the inferred clusters, indicated with the color bar on top.



**Fig 7.** Enrichment analysis results for each set of DMGs.

# Discussion

As Fig. 1 shows, the vast majority of cells in each sample are neutrophils. Furthermore, a considerable number of samples have a significant number of T helper cells and monocytes. On the other hand, cell types like B cells, eosinophils and natural killer cells (NK cells) are negligible in most analyzed samples. It must be noted too that meth-atlas detected small percentages of other cell types that were not present in the reference data.

Concerning samples clustering, Fig. 2 strongly suggests that there are 3 clusters and a possible outlier. This is further backed by the PCA in Fig. 3, where the samples seem to group according to the inferred clusters. Moreover, the sample colored in black (assumed to be outlier) in the PCA, is far from the others, which reinforces it status as an outlier. Furthermore, these clusters tend to correlate quite well with the methylation patterns across samples (Fig. 4). All in all, these analyses together point in the same direction, and thus the clusters were used for further analysis.

On the other hand, the correlation between DMGs and CpG clusters (Fig. 5) suggests that there is not much difference between each set of DMGs that was obtained (in other words, they share many CpGs in common). Although an interesting pattern that can be observed is that CpGs that were not differentially methylated in any comparison, are mostly those assigned to the cluster represented with color green (most of them, CpGs with very low methylation levels). All of this makes more sense when Fig. 6 is taken into consideration: while the red and blue clusters seem to be clearly defined and their members strongly correlated, the green cluster is sparser, which indicates that those CpGs do not have a clear methylation pattern across samples (which might explain why the majority of them are non-differentially methylated between clusters).

Finally, enrichment analyses further seem to suggest that the 3 sets of DMGs obtained from each cluster comparison, share most of CpGs. this is because enriched terms in every case are very similar. They consist mostly of immune response related terms, and more specifically, T cell response. This result is particularly intriguing, considering that the vast majority of cells, according to the deconvolution analysis (Fig. 1) are neutrophils. However, T cells (CD4 and CD8 together) conform overall the second largest group of cells in most samples. Further research is need to determine whether the sample clusters that were inferred here correlate with the number of T cells in the sample. That would explain the predominance of T cells related terms in the enrichment analysis.