Appendix I: Clustering

Marthe Veldhuis

At the first few weeks of the project, I tried using ClusterExperiment package, in combination with SummarizedExperiment to explore the data initially.

Before we had any proper questions defined, I tried to cluster all data initially. This did insinuate some sort of grouping as can be seen in the image below. However, it was difficult to relate it to anything but the top genes that differentiated the most between the groups.

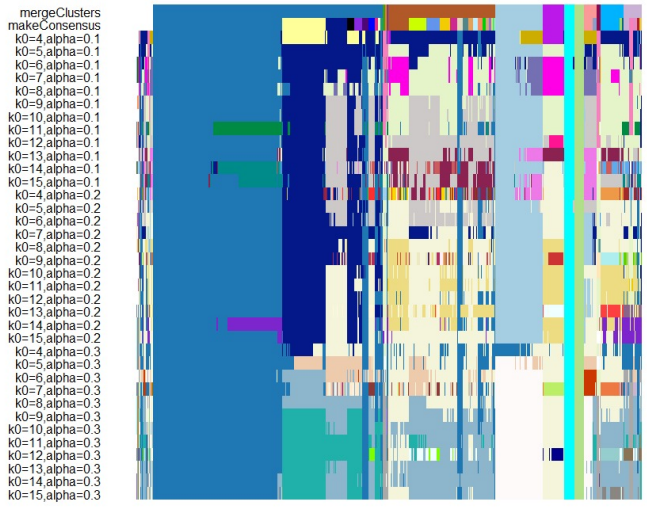

Figure 1: Clustering of all data

Therefore I tried to cut down the sample size by focusing on good responders only, and then add the results of a meta data variable alongside the results as can be seen below.

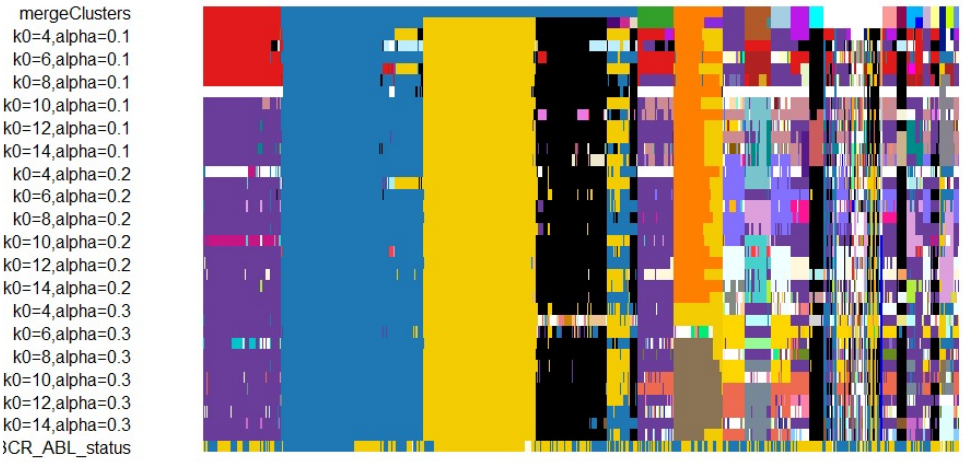

Figure 2: Clustering of good responder data. BCR_ABL_status added

After we defined the questions, I combined these approaches to the specific goal of distinguishing good and poor responders. Therefore clustering only these patients at diagnosis time, and adding the responder status to the result graph.

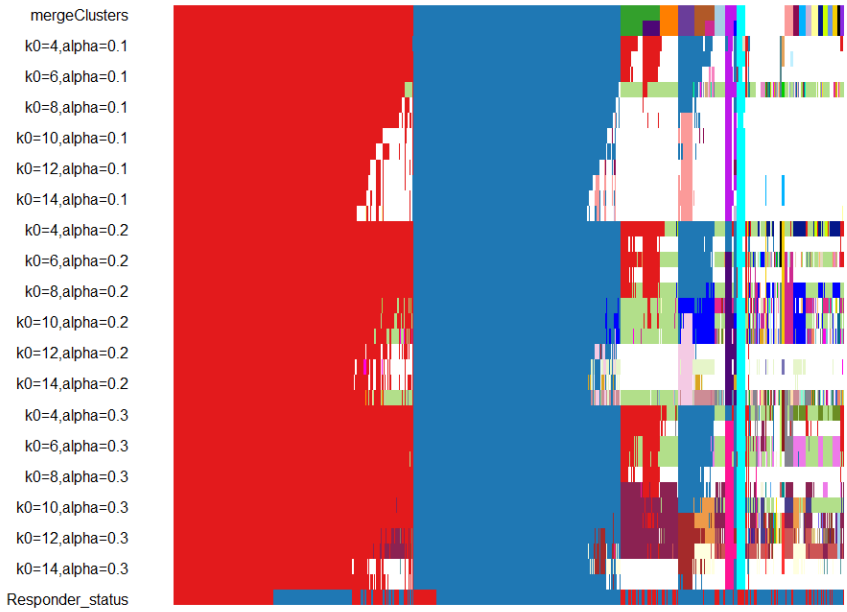

Figure 3: Clustering of good and poor responders data only. Responder_status added

This definitely showed promising results, which is why I wanted to continue with this type of work. However, since this package (ClusterExperiment) did not allow the meta variables to be included in the process of the clustering itself (they can only be shown in the results), I dropped this software. Instead I looked at the underlying processes, causing me to find the limma package.