Supplementary Material

Using repeated measurements to validate hierarchical gene clusters

December 11, 2007

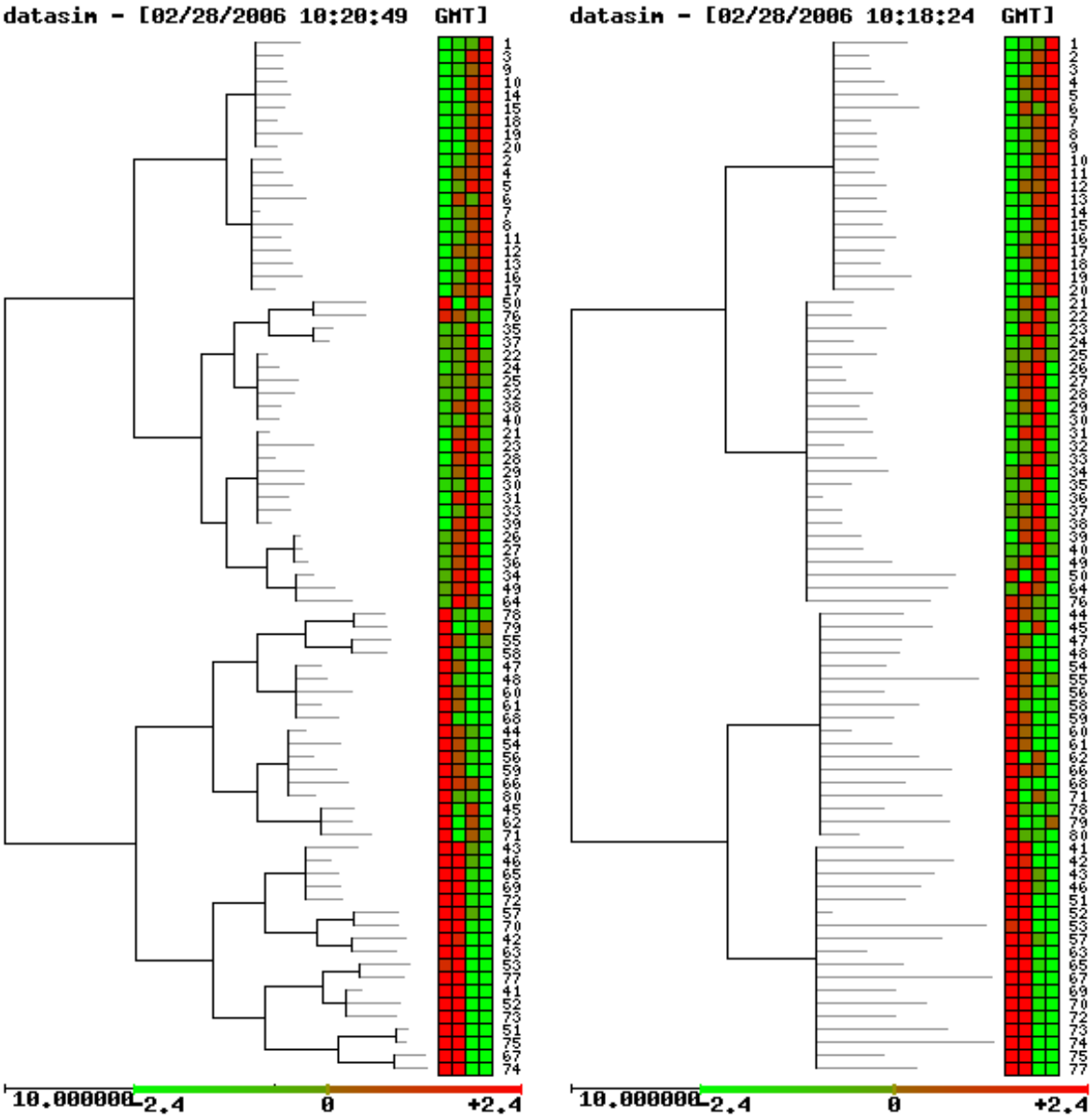
In this Supplementary Material, additional analyses concerning the paper entitled "Using repeated measurements to validate hierarchical gene clusters" are provided. Section 1 presents the clustering achieved with the SOTA software when analysing the simulated data of Numerical Experiments section. Section 2 presents the results achieved on the wood data set when using the phylogenetic bootstrap proportion, instead of the measure proposed in Expression (1). Section 3 presents a detailed analysis of the transcriptomic study on iron stress for *Arabidopsis Thaliana*.

1 Results of SOTA with simulated data

Statistical units of cluster 1 are labeled from 1 to 20, those of cluster 2 from 21 to 40, and those of cluster 3 from 41 to 80. Clustering is achieved using the Euclidean distance. Figure 1-a gives the hierarchical clustering with *Variability threshold* equal to 90% (default value) and Figure 1-b the clustering with *Variability threshold* equal to 30%. As discussed in the article, whatever the threshold, it is impossible to recover the simulated clusters.

2 Results using phylogenetic bootstrap proportions

Figure 2 presents the results achieved on the wood data set using the bootstrap proportion, instead of the Jaccard index of Expression (1). The bootstrap proportion (BP) is widely used in phylogenetics; it counts the percentage of cases where the studied cluster is exactly recovered by the pseudo-trees. All BP values above 0.1 are shown in Figure 2. Apart from some very small clusters, this approach is unable to highlight interesting clusters (compared with Figure 2 in the article), due to the fact that large clusters are (almost) never exactly recovered. Figure 3 compares the distributions among clusters of BP and Jaccard index.



(a) (b)

Figure 1: Analysis of simulated dataset with SOTA.

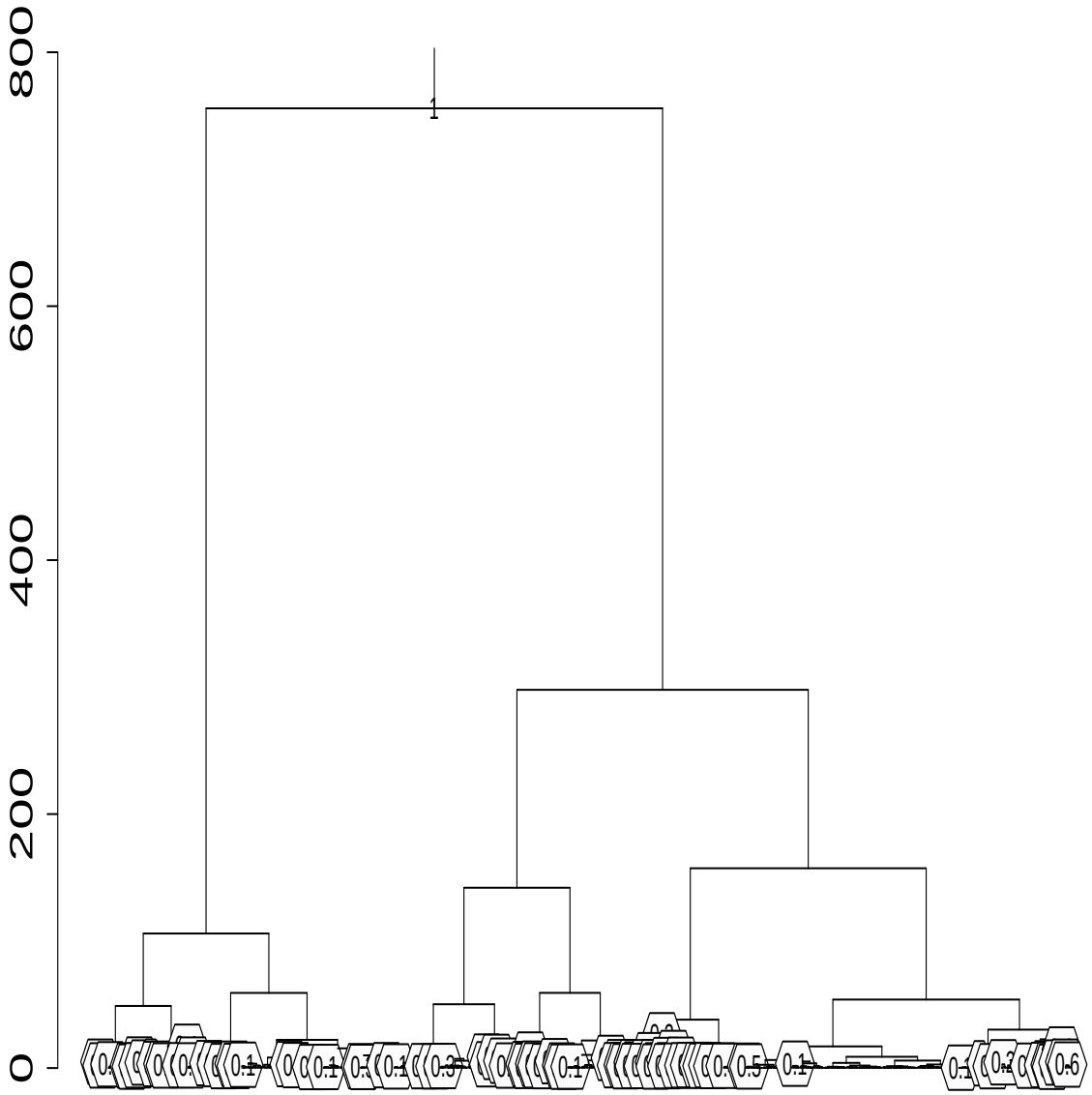
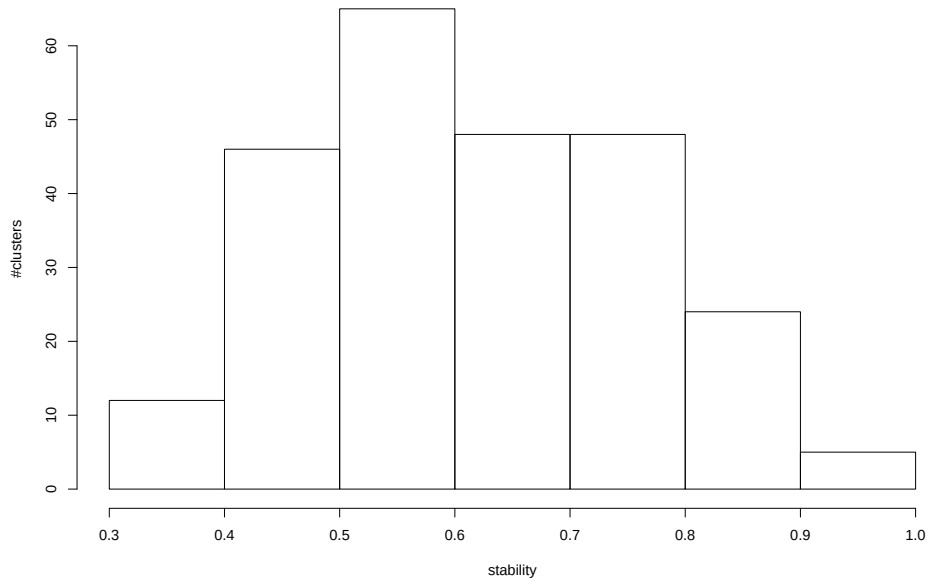
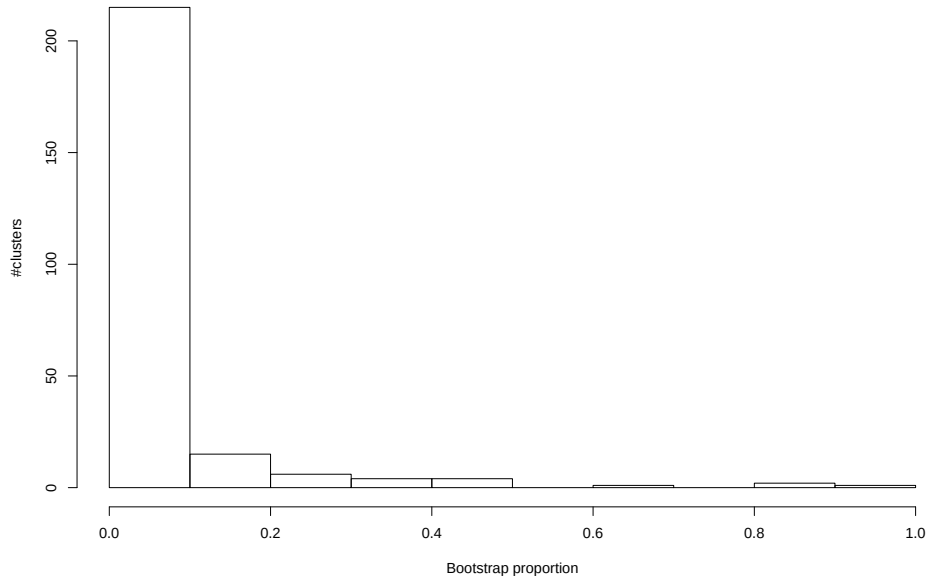


Figure 2: Results achieved on the wood data set, when using the phylogenetic bootstrap proportion (BP). All nodes with BP above 0.1 are indicated.



(a)



(b)

Figure 3: Distribution of the Jaccard index (a) and of the phylogenetic bootstrap proportion (b), among clusters of size > 5.

3 Transcriptomic study on iron stress for *A. Thaliana*

Figure 4 shows a detailed view of Cluster C presented in Figure 5-a of the article. As discussed in the paper, this cluster has good stability (about 0.8) and exhibits several over-represented GO terms. When detailing the composition of this cluster, it appears that only one of its sub-clusters has high stability (around 0.77), though the other sub-clusters have stabilities below 0.7. Thus, while the large cluster C has high stability and constitutes a well defined class, its inside organisation is not so stable: only the 0.77 cluster seems to define a real, well defined inside class.

We used the GOStat[1] tools of Beissbarth and Speed [1] on the different sub-clusters. Again, only the stable sub-cluster exhibited over-represented GO terms. Table 1 and 2 give the over-represented GO terms in the large 0.8 cluster and the small 0.77 sub-cluster, respectively, along with the gene lists annotated with these terms. All over-represented annotations involve stress response, which closely corresponds to the experiment conducted in this study. Genes in the 0.77 sub-cluster are mainly involved in heat response (7 among 12), while genes in the 0.8 cluster are also involved in response to light. Among the genes that are in the 0.8 cluster but not in the 0.77 sub-cluster, 8 are involved in response to heat, and 4 (among a total of 5) in response to light. Hence, all the biological information of the 0.8 cluster is not solely concentrated in the 0.77 sub-cluster, but also involves several other genes and functions.

References

[1] T. Beissbarth and T. P. Speed. Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9), 2004.

[1] `http://gostat.wehi.edu.au/`

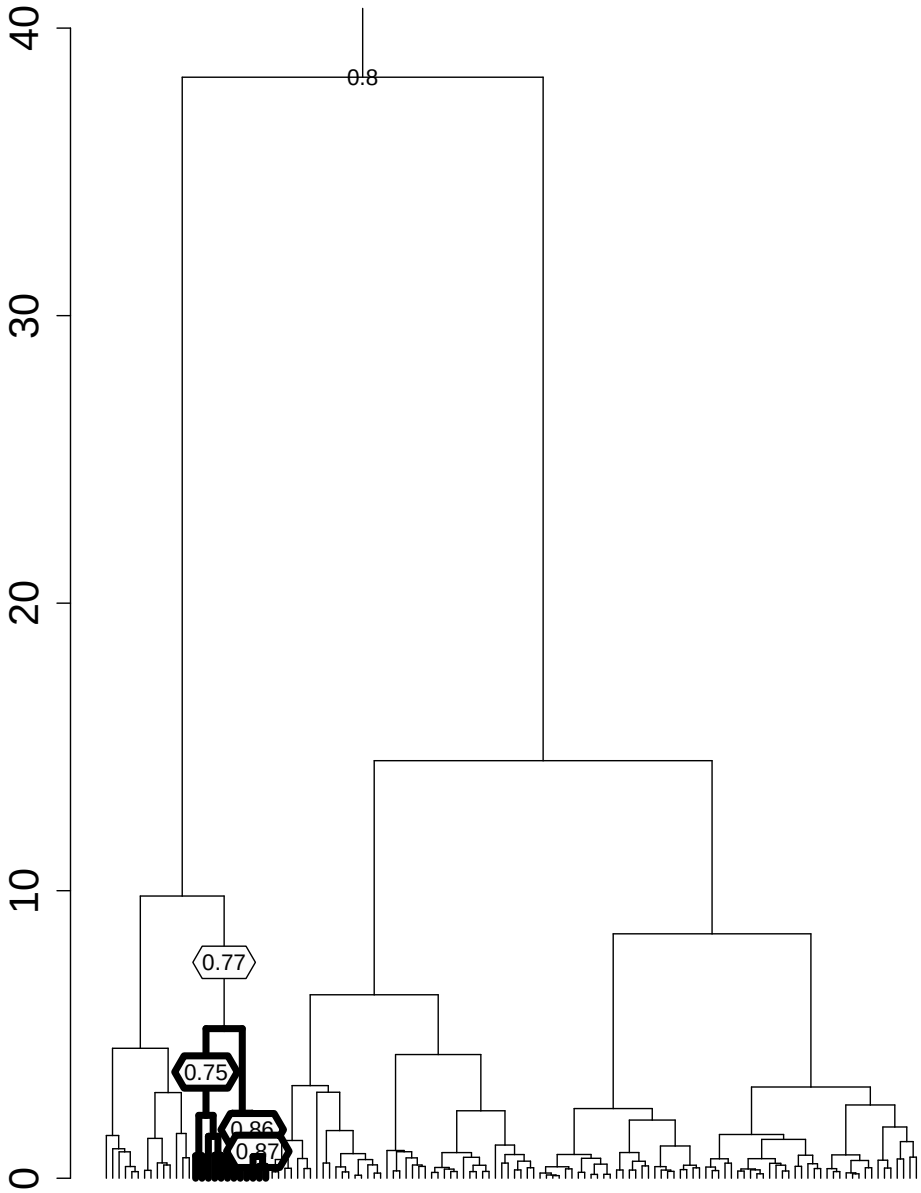


Figure 4: A zoom of cluster C (see article). All clusters with stability > 0.7 are shown.

GO term id	GO term name	gene list	p-value
GO:0009408	response to heat	MBF1C AT1G59860.1 AT3G51910.1 HSA32 AT5G51440.1 AT5G02490.1 AT4G25200.1 AT1G07400.1 ATHSP17.4 HSP70B AT2G19310.1 AT5G37670.1 ATBAG6 AT2G29500.1 ATGOLS1	1.13e-08
GO:0009628	response to abiotic stimulus	MBF1C AT1G59860.1 AT3G51910.1 HSA32 HSP70B AT2G19310.1 ELIP1 ATBAG6 ATGOLS1 AT5G02490.1 AT5G51440.1 OST1 AT4G25200.1 AT1G07400.1 ATHSP17.4 AT5G37670.1 AT2G29500.1 AT5G58070.1	6e-08
GO:0009266	response to temperature stimulus	MBF1C AT1G59860.1 AT3G51910.1 HSA32 AT5G02490.1 AT5G51440.1 AT4G25200.1 ATHSP17.4 AT1G07400.1 HSP70B AT2G19310.1 ELIP1 AT5G37670.1 ATBAG6 AT2G29500.1 ATGOLS1 AT5G58070.1	2.62e-07
GO:0006950	response to stress	MBF1C AT1G59860.1 AT3G51910.1 HSA32 HSP70B AT2G19310.1 ELIP1 ATBAG6 ATGOLS1 AT5G02490.1 AT5G51440.1 OST1 AT4G25200.1 AT1G07400.1 ATHSP17.4 AT5G37670.1 AT2G29500.1 AT5G58070.1	0.000936
GO:0009644	response to high light intensity	AT2G19310.1 AT3G51910.1 ATBAG6 AT2G29500.1 ATGOLS1	0.0091
GO:0009642	response to light intensity	AT2G19310.1 AT3G51910.1 ATBAG6 AT2G29500.1 ATGOLS1	0.0091
GO:0050896	response to stimulus	MBF1C AT1G59860.1 AT3G51910.1 HSA32 HSP70B AT3G28210.1 AT2G19310.1 ELIP1 ATBAG6 ATGOLS1 AT5G02490.1 AT5G51440.1 OST1 AT4G25200.1 ATHSP17.4 AT1G07400.1 ATERF13 AT5G37670.1 AT2G29500.1 AT1G63840.1 AT5G58070.1	0.00928

Table 1: Over-represented GO terms in cluster with stability 0.8.

GO term id	GO term name	gene list			p-value
GO:0009408	response to heat	MBF1C	AT1G59860.1	AT5G51440.1 AT2G29500.1 AT4G25200.1 AT1G07400.1 ATHSP17.4	5.39e-08
GO:0009266	response to temperature stimulus	MBF1C	AT1G59860.1	AT5G51440.1 AT2G29500.1 AT4G25200.1 AT1G07400.1 ATHSP17.4	5.62e-07
GO:0009628	response to abiotic stimulus	MBF1C	AT1G59860.1	AT5G51440.1 AT2G29500.1 AT4G25200.1 AT1G07400.1 ATHSP17.4	1.77e-06
GO:0006950	response to stress	MBF1C	AT1G59860.1	AT5G51440.1 AT2G29500.1 AT4G25200.1 AT1G07400.1 ATHSP17.4	1.99e-05
GO:0050896	response to stimulus	MBF1C	AT1G59860.1	AT5G51440.1 AT2G29500.1 AT4G25200.1 AT1G07400.1 ATHSP17.4	0.000211

Table 2: Over-represented GO terms in cluster with stability 0.77.