Supplementary material: A Newton-based variant of Exclusive Lasso for improved sparse solutions

Dayasri Ravi[1*] and Andreas Groll[1]

[1]Department of Statistics, TU Dortmund University, Vogelpothsweg 87, Dortmund, 44221, NRW, Germany.

*Corresponding author(s). E-mail(s): ravi@statistik.tu-dortmund.de;
Contributing authors: groll@statistik.tu-dortmund.de;

Weighted gene co-expression networks

Weighted gene co-expression networks (WGCNA) are widely used in systems biology to investigate gene functionality on a system-wide scale. We used this approach to group genes into modules based on high topological overlap, where genes with highly correlated expression profiles across samples were clustered together. We considered these modules identified by WGCNA as groups for our Lasso problems. The methodology began with preprocessing the cancer gene expression data to ensure it was clean and normalized for analysis. The dataset was then divided into two parts: one consisting of tumor-sensitive samples and the other of normal samples. We used the tumor data to construct the gene co-expression network. First, we determined the appropriate soft-thresholding power for the network topology. Using a range of powers (1-50), we applied the `pickSoftThreshold` function from the `WGCNA` R package (Langfelder and Horvath, 2008). This function analyzed the scale-free topology fit for different powers. We then plotted the scale-free topology model fit and mean connectivity against the power values. From these plots, we identified a suitable soft-thresholding power.

With the chosen power, we constructed the weighted gene co-expression network using the `blockwiseModules` function from `WGCNA` package. This function identified modules by clustering genes based on their topological overlap. We set an appropriate maximum block size, used a signed network type, and merged modules with a cut height of 0.25. The analysis yielded module **eigengenes**, which represented the first principal component of each module, summarizing the expression profiles of the genes within the module.

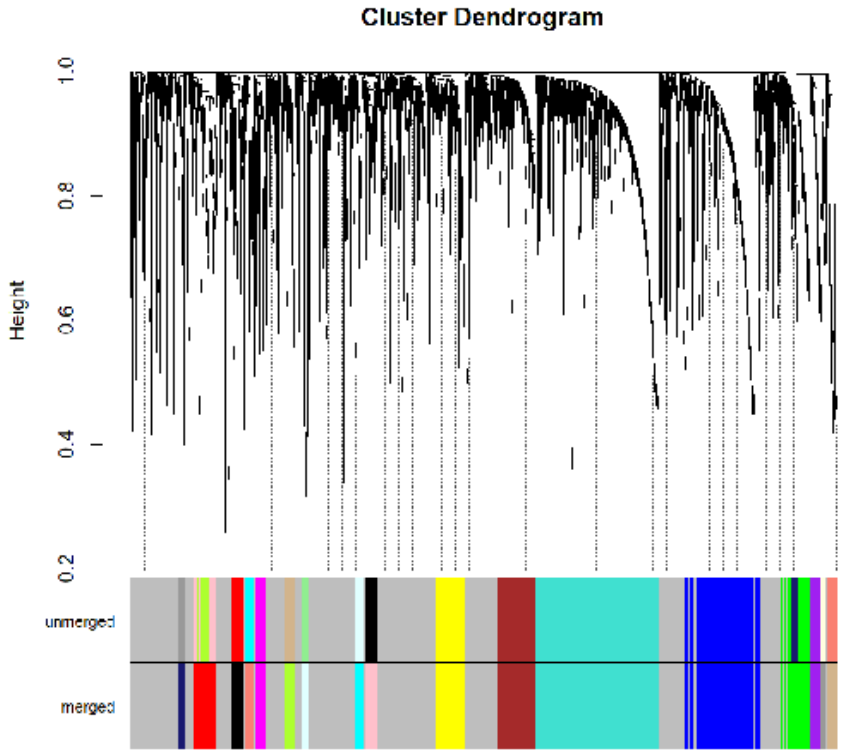

Figure S1: Identification of network modules in OC dataset 1

We examined the number of genes in each module and visualized the dendrogram of gene clustering alongside the module colors, both before and after merging. Finally, these identified modules were used to categorize genes into groups for both tumor and normal samples.

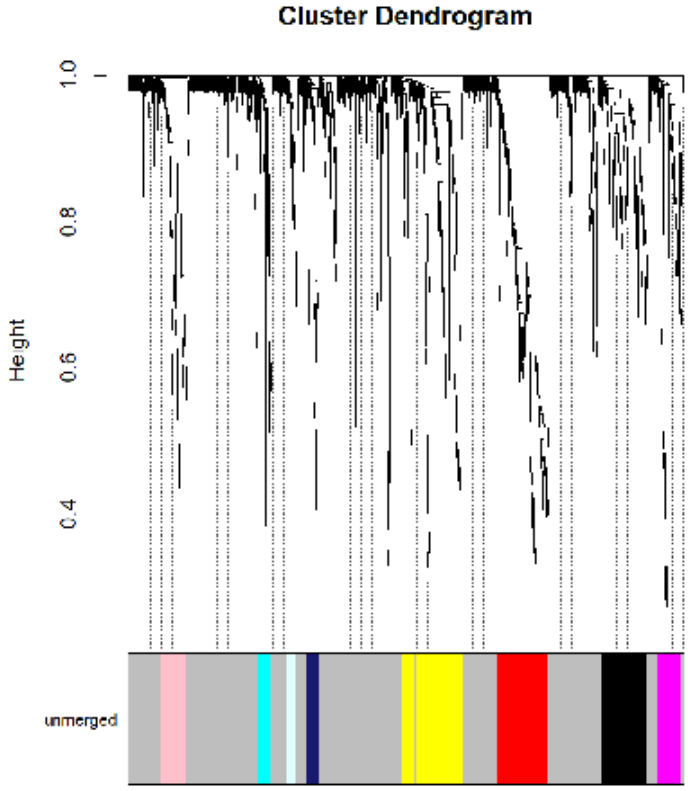

Figure S2: Identification of network modules in OC dataset 2

Figures S1 and S2 show the Cluster Dendrogram of OC dataset 1 and OC dataset 2, respectively. We chose a maximum block size of 3,000 for our analysis. However, because our dataset contained more than 3,000 genes, the figure illustrates the results for only one block as an example.

We excluded unmerged genes labeled as grey that do not belong to any module. The number of genes in each group is shown in Tables S1 and S2 for OC datasets 1 and 2, respectively.

References

Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics **9**, 1–13 (2008)

Module	Number of Genes
black	62
blue	335
brown	181
cyan	33
green	130
greenyellow	47
lightcyan	28
magenta	51
midnightblue	29
pink	56
purple	51
red	106
salmon	42
tan	45
turquoise	569
yellow	134

Table S1: OC dataset 1: Number of genes in each module

Module	Number of Genes
black	112
blue	450
brown	200
cyan	31
green	135
greenyellow	55
lightcyan	26
lightgreen	24
lightyellow	24
magenta	61
midnightblue	31
pink	66
purple	58
red	129
salmon	44
tan	47
turquoise	3894
yellow	144

Table S2: OC dataset 2: Number of genes in each module