

Supplementary Figures:

Supplementary Figure 1 – PANCAN Analysis

Supplementary Figure 2 – Overexpressed transcripts and enriched gene ontology terms in subgroup 1a

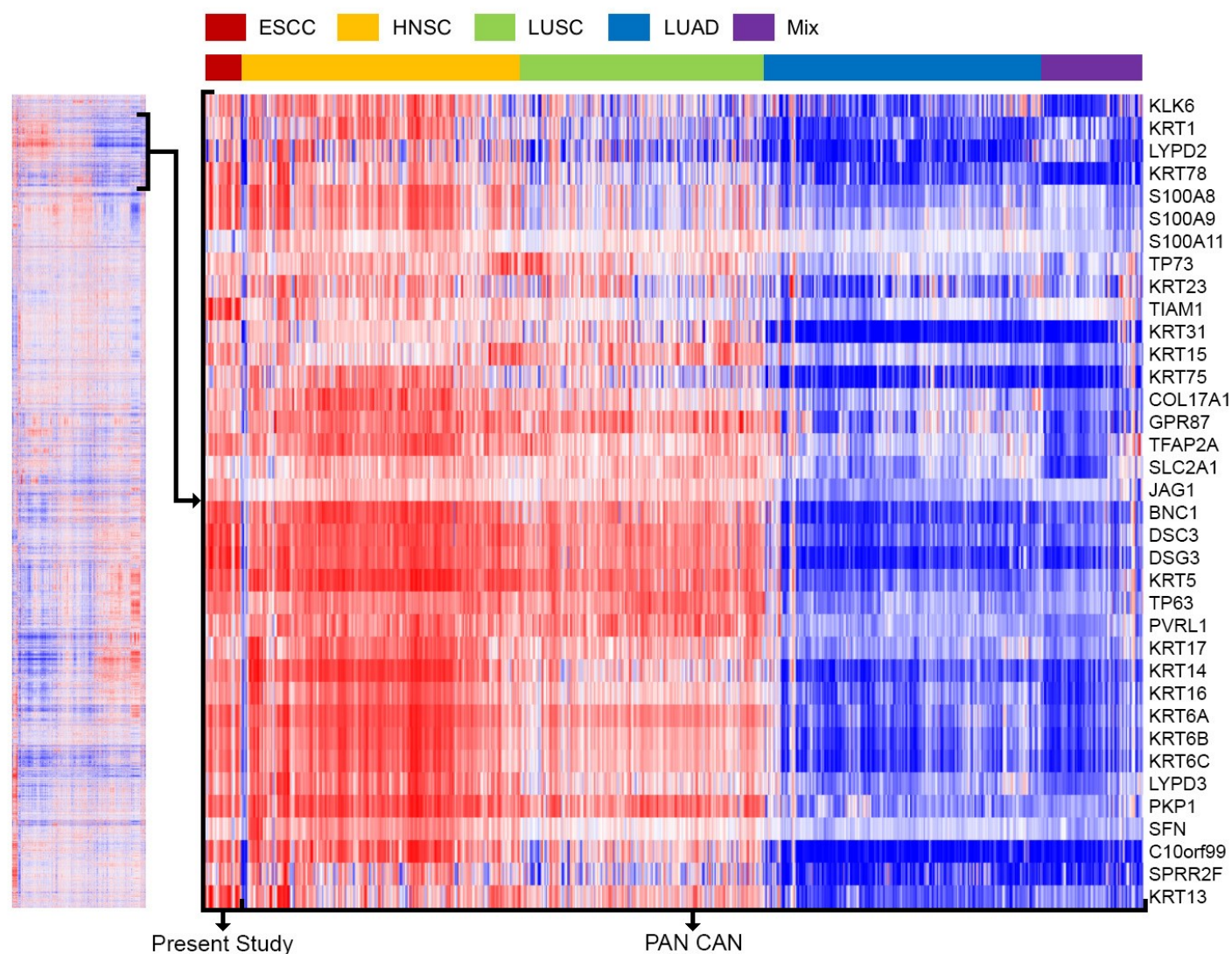
Supplementary Figure 3 – Overexpressed transcripts and enriched gene ontology terms in subgroup 1b

Supplementary Figure 4 – Underexpressed transcripts and enriched gene ontology terms in subgroup 1b

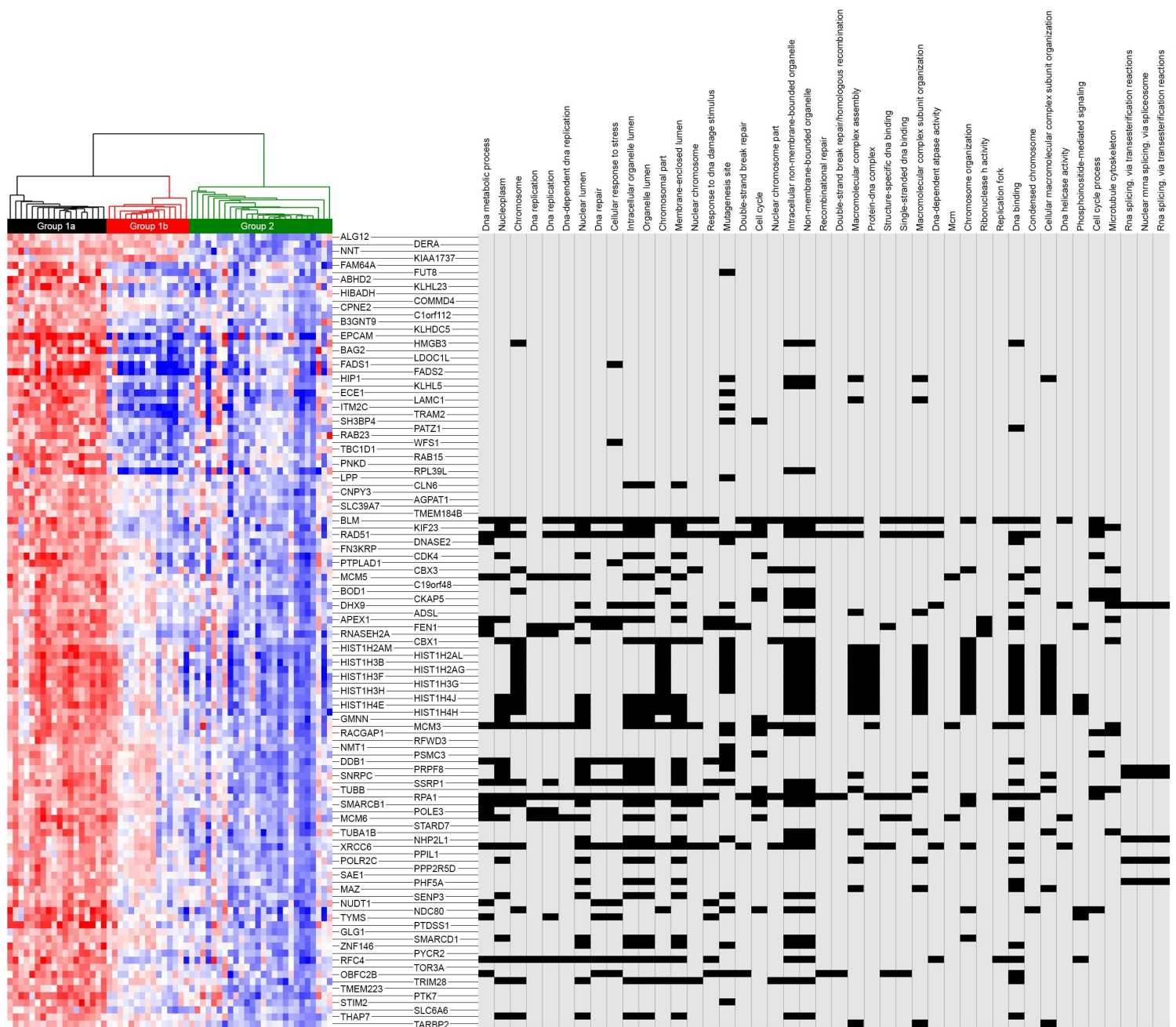
Supplementary Figure 5 – Overexpressed transcripts and enriched gene ontology terms in subtype 2

Supplementary Figure 6 – RNA subtype clustering silhouette widths

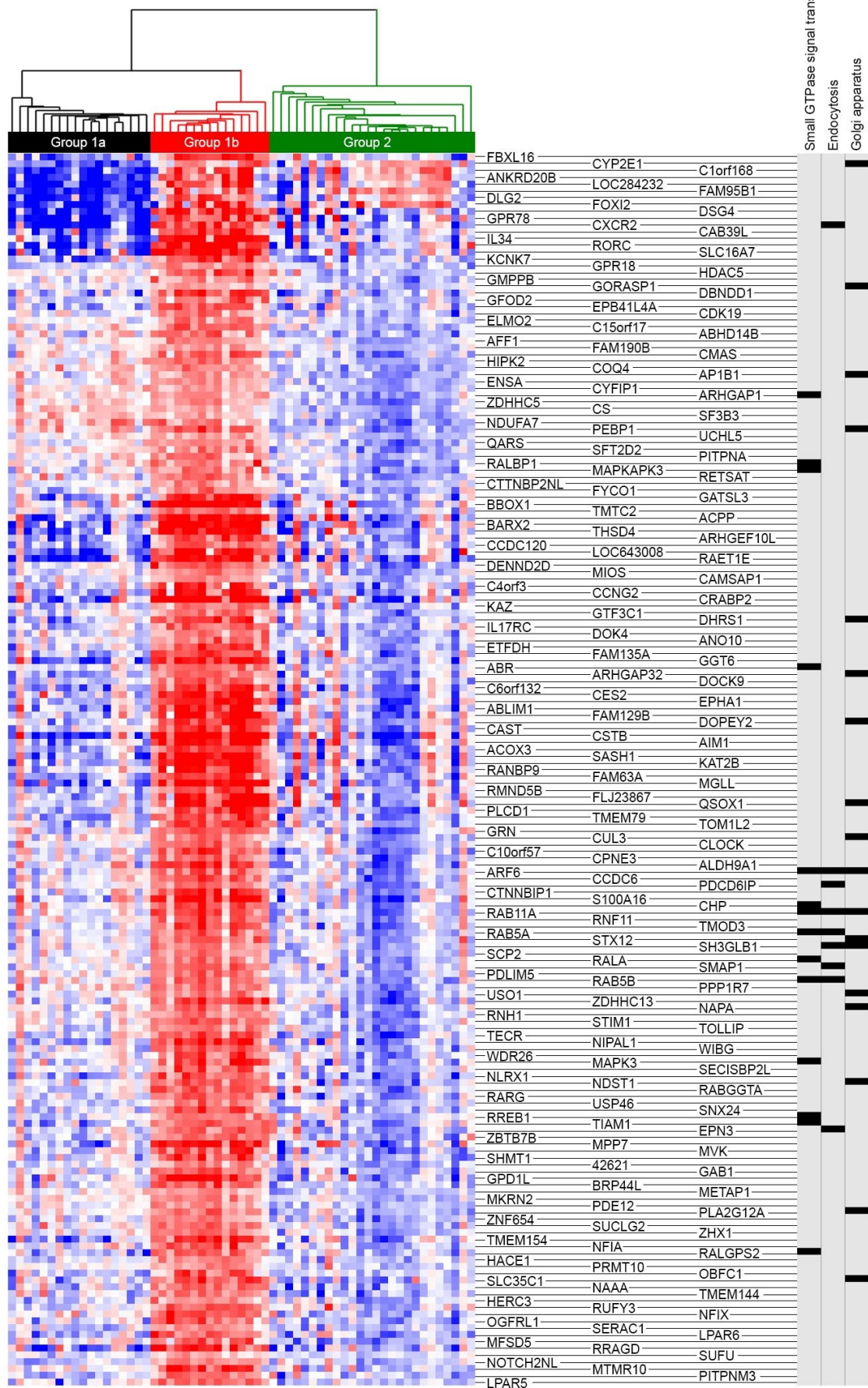
Supplementary Figure 7 – RNA subtype clustering principal component analysis



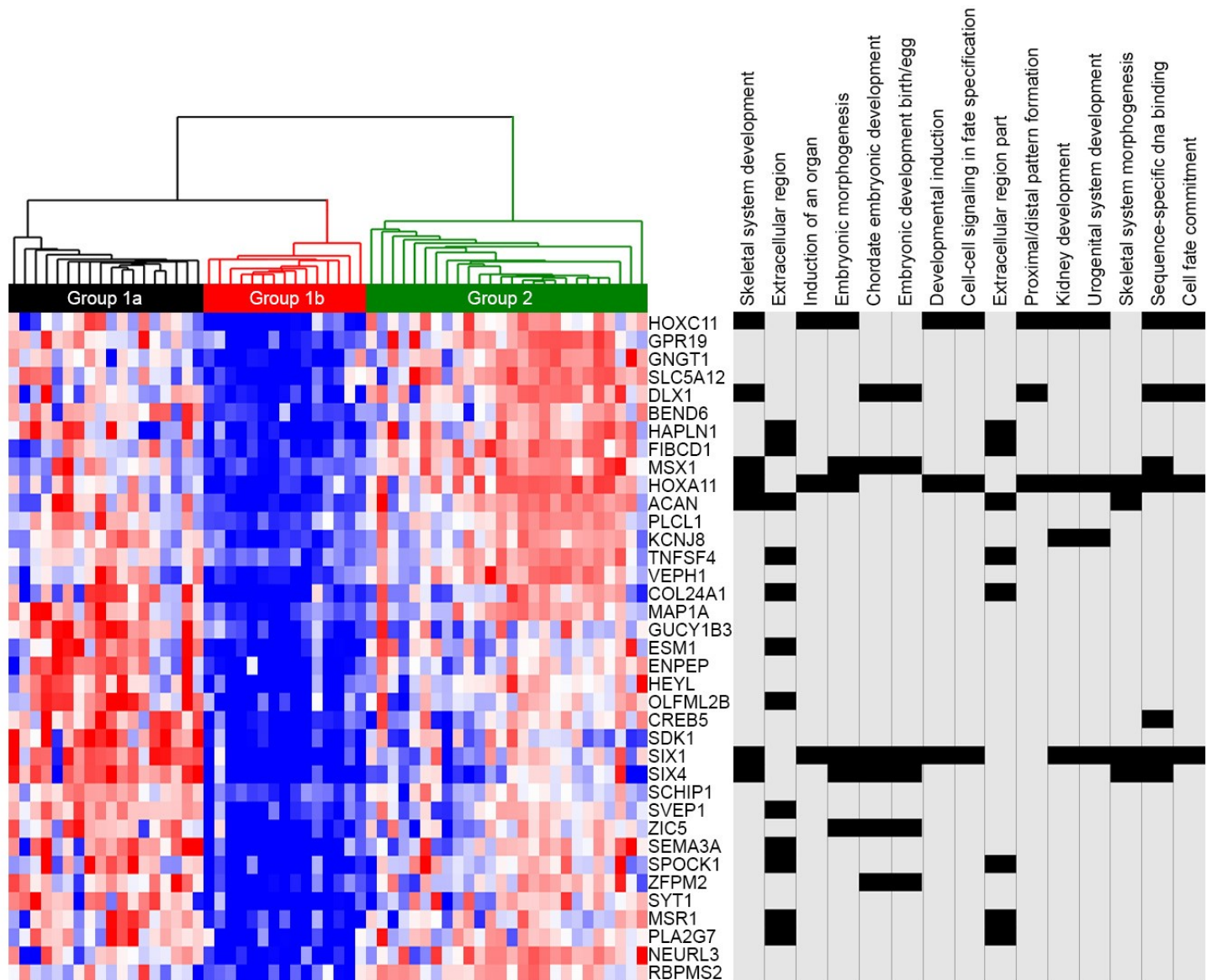
Supplementary Figure 1 – PANCAN Analysis. Malawian ESCC RNAseq samples were clustered against TCGA data sets from the PANCAN analysis (30). ESCC samples clustered together compared to other PANCAN tumor types and nearest to squamous carcinomas of the head and neck and lung.



Supplementary Figure 2 – Overexpressed transcripts and enriched gene ontology terms in subgroup 1a. The heatmap on the left of the figure demonstrates transcripts dynamically overexpressed (red) in subgroup 1a, where each row is a single transcript and each column is a single sample. The data matrix on the right of the figure demonstrates enriched gene ontology (GO) terms, where each row is a single transcript and each column is a single GO term, ordered left-to-right by increasing false discovery rate. The transcript names displayed in the middle of the figure are aligned horizontally to the corresponding row of the expression heatmap on the left and the GO matrix on the right.

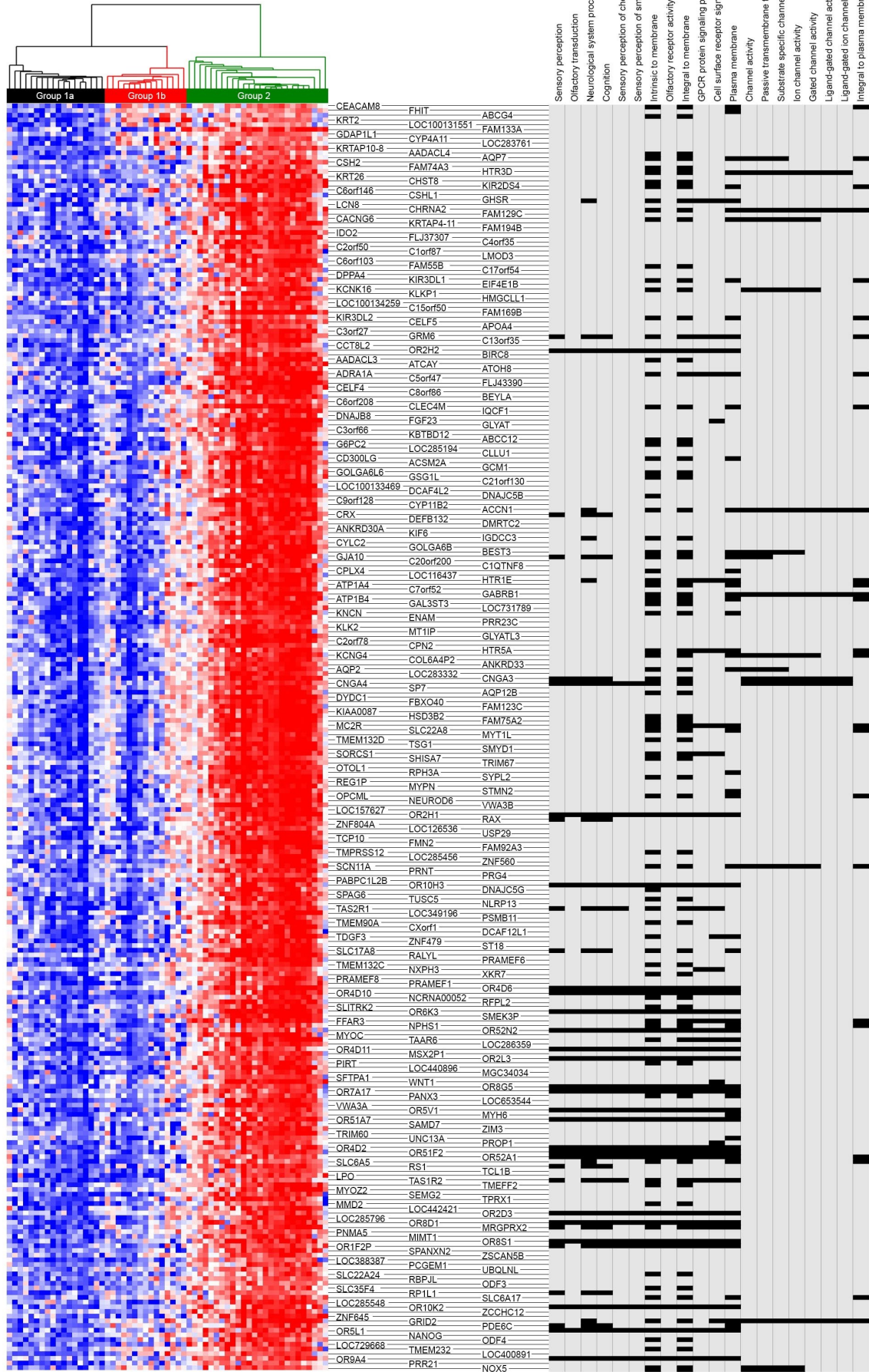


Supplementary Figure 3 – Overexpressed transcripts and enriched gene ontology terms in subgroup 1b. The heatmap on the left of the figure demonstrates transcripts dynamically overexpressed (red) in subgroup 1b, where each row is a single transcript and each column is a single sample. The data matrix on the right of the figure demonstrates enriched gene ontology (GO) terms, where each row is a single transcript and each column is a single GO term, ordered left-to-right by increasing false discovery rate. The transcript names displayed in the middle of the figure are aligned horizontally to the corresponding row of the expression heatmap on the left and the GO matrix on the right.



Supplementary Figure 4 – Underexpressed transcripts and enriched gene ontology terms in subgroup 1b.

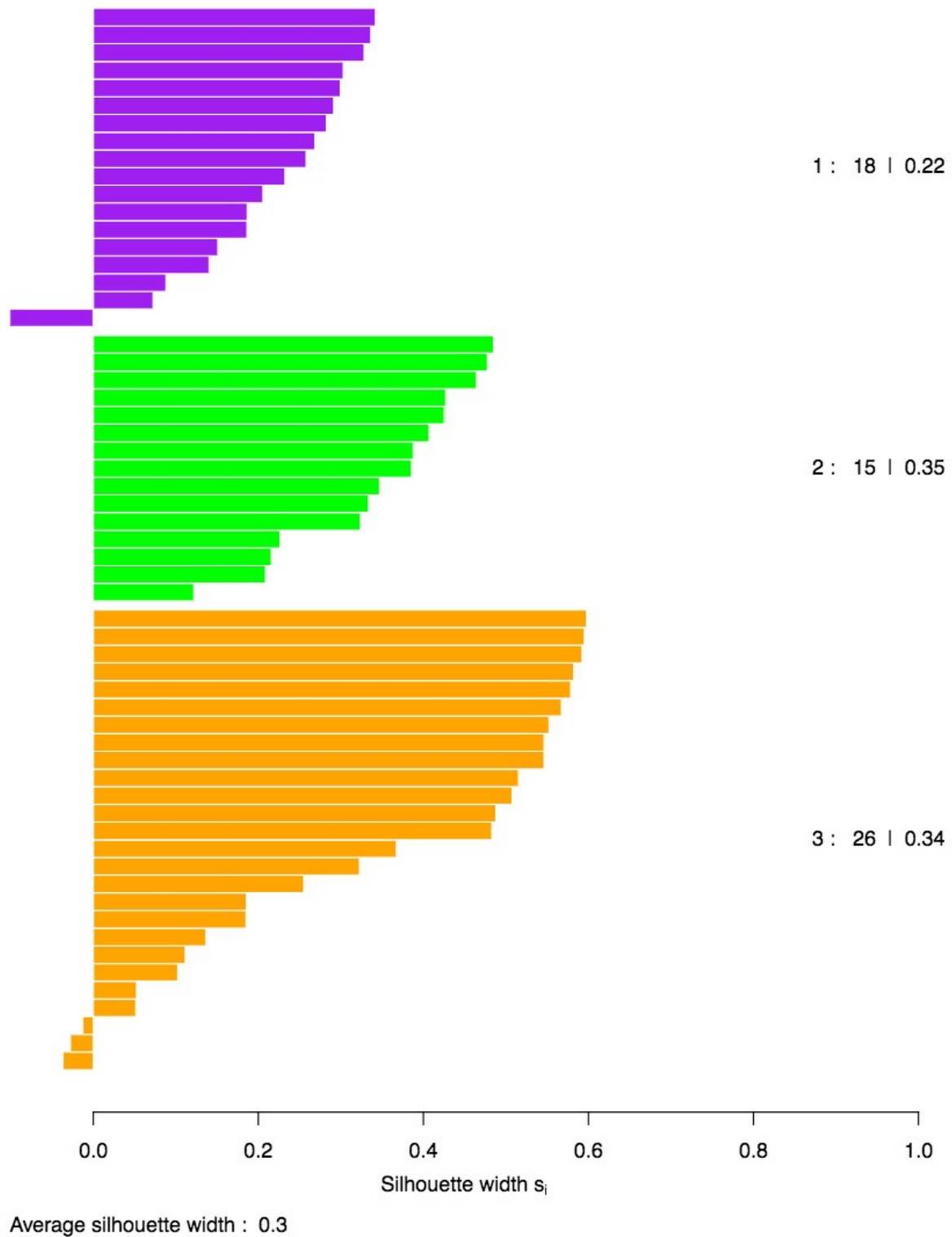
The heatmap on the left of the figure demonstrates transcripts dynamically underexpressed (blue) in subgroup 1b, where each row is a single transcript and each column is a single sample. The data matrix on the right of the figure demonstrates enriched gene ontology (GO) terms, where each row is a single transcript and each column is a single GO term, ordered left-to-right by increasing false discovery rate. The transcript names displayed in the middle of the figure are aligned horizontally to the corresponding row of the expression heatmap on the left and the GO matrix on the right.



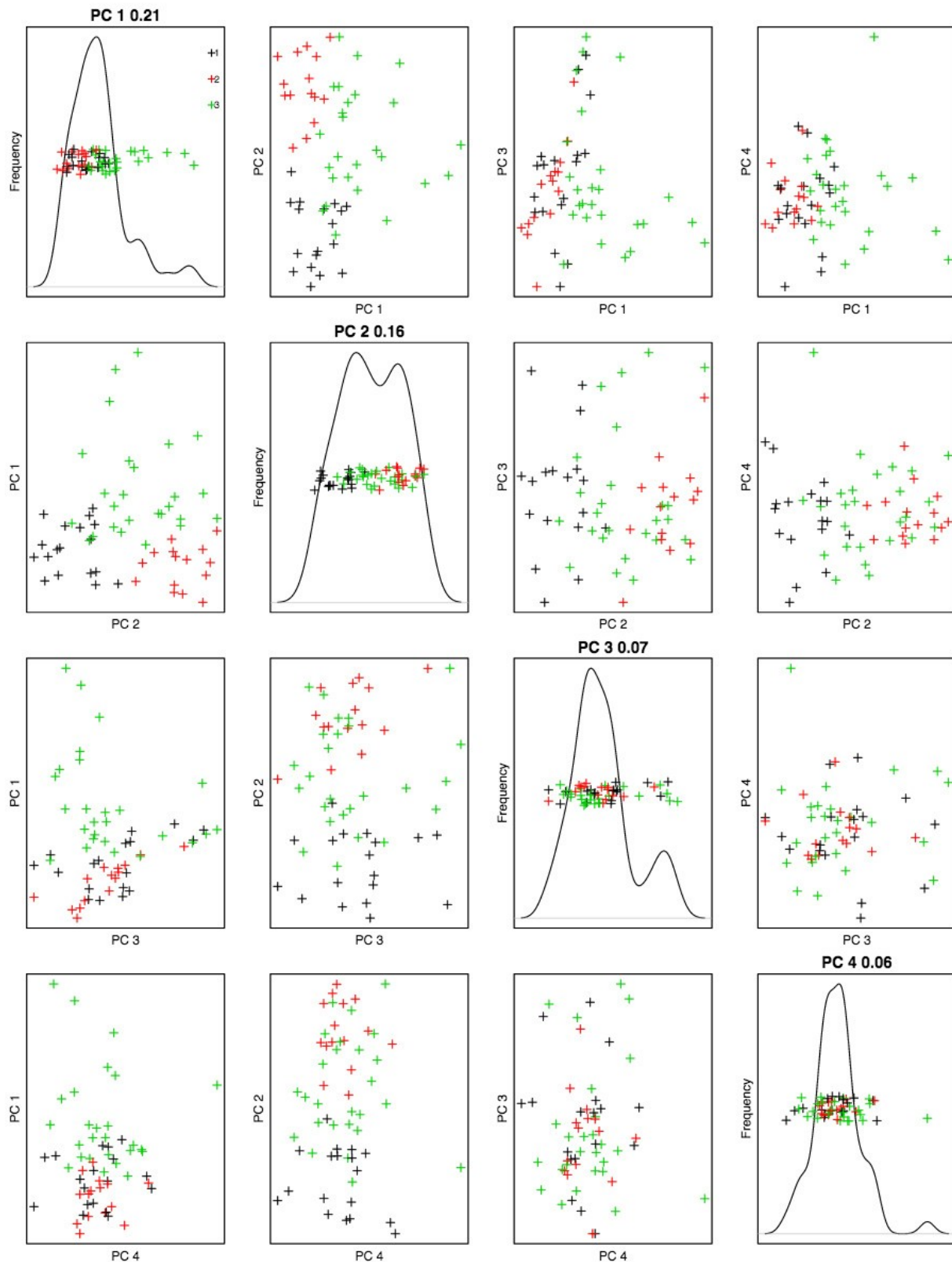
Supplementary Figure 5 – Overexpressed transcripts and enriched gene ontology terms in subtype 2. The heatmap on the left of the figure demonstrates transcripts dynamically overexpressed (red) in subtype 2, where each row is a single transcript and each column is a single sample. The data matrix on the right of the figure demonstrates enriched gene ontology (GO) terms, where each row is a single transcript and each column is a single GO term, ordered left-to-right by increasing false discovery rate. The transcript names displayed in the middle of the figure are aligned horizontally to the corresponding row of the expression heatmap on the left and the GO matrix on the right.

n = 59

3 clusters C_j
 $j : n_j \mid \text{ave}_{i \in C_j} s_i$



Supplementary Figure 6 – RNA subtype clustering silhouette widths. Silhouette width is a measure of similarity between a sample and its cluster, compared to other clusters. The average silhouette width of 0.3 and predominantly non-negative values in each cluster support three ESCC subtypes. Purple represents subgroup 1a, green represents subgroup 1b, and orange represents subtype 2.



Supplementary Figure 7 – RNA subtype clustering principal component analysis. The two most dominant principal components separate the clusters and support three ESCC subtypes. Principal component 1 (PC1) predominantly separates subtype 1 (black and red) and subtype 2 (green). Principal component 2 (PC2) predominantly separates subgroup 1a (black) and subgroup 1b (red).