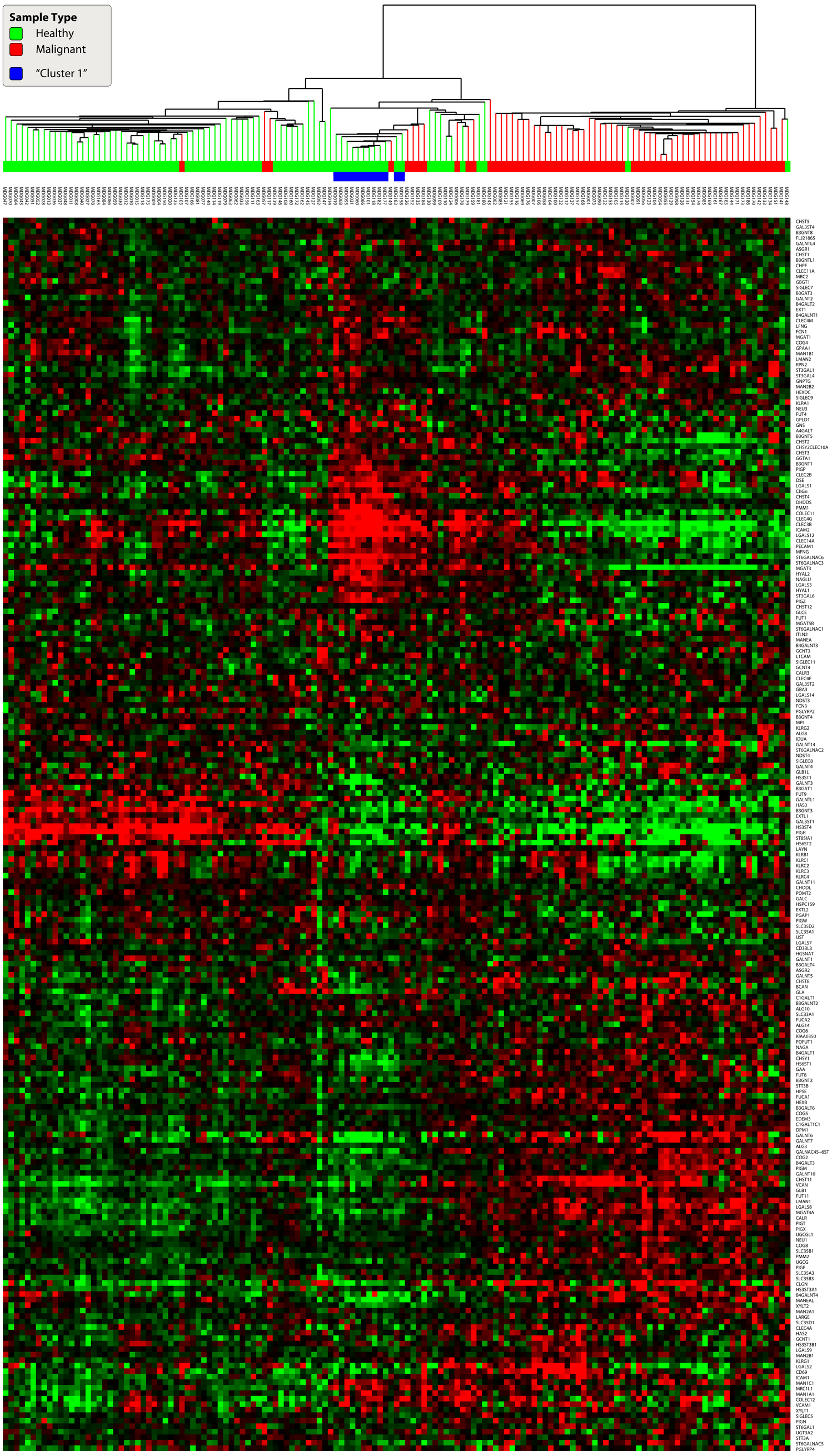
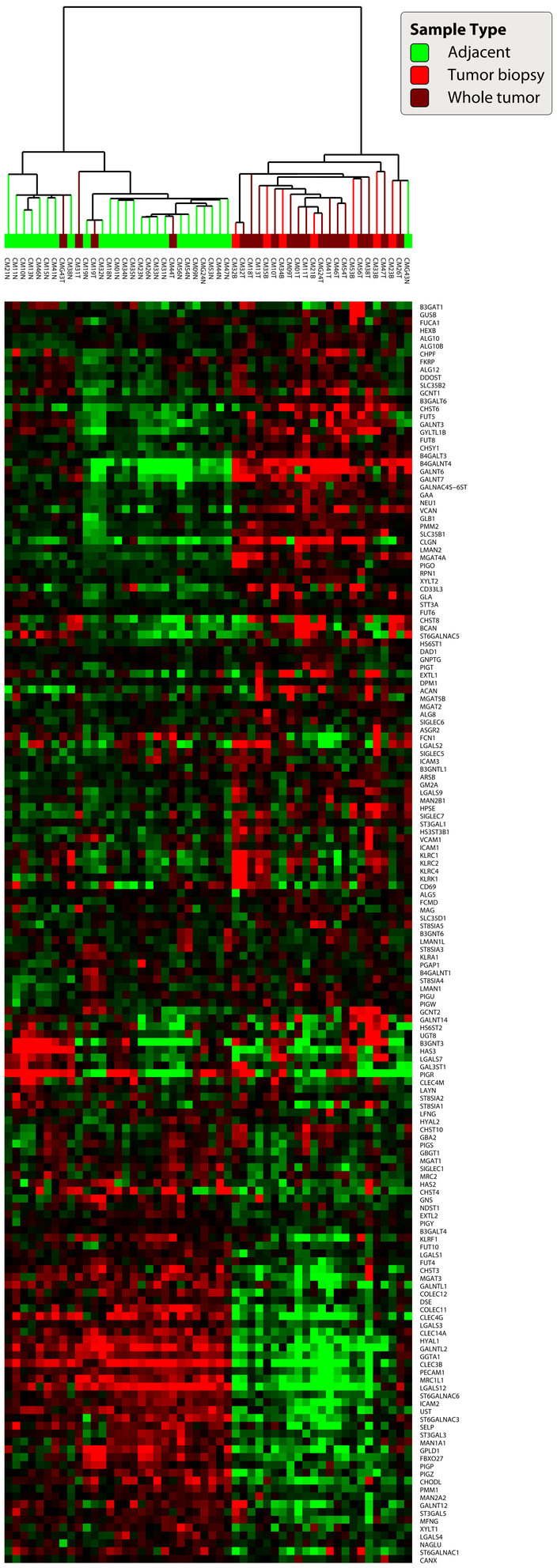
Supplementary figure 1: Hierarchical clustering of glycan gene expression including sample names and gene symbols.

A) *Dataset A*. Malignant tissue and samples from healthy women are marked red and green on the dendrogram, respectively. These two groups are well separated indicating that glycan gene expression is profoundly altered during carcinogenesis. A subgroup of normal samples appears to have a different glycan gene signature. This group is largely identical to the "cluster 1" group defined by whole-genome expression analysis by Haakensen et al (unpublished data). Samples classified as "cluster 1" are marked blue.

B) *Dataset B*. Tumor tissue samples are marked red on the dendrogram – bright red if extracted from a tumor biopsy or dark red in case whole-tumor was used. Core biopsies of adjacent breast tissue are marked green. The malignant and the non-malignant samples are well separated in this dataset as well. Note that the normal and malignant samples are equally well separated in both datasets (with only a few misplaced samples) despite that dataset A contains samples from different individuals while dataset B consists of tumor and adjacent normal tissue from the same patient.

a)



b)