Supplemental Data

Proteins with evolutionarily hypervariable domains are associated with immune response and better survival of basal-like breast cancer patients

Shutan Xu, Shaying Zhao

Inventory of Supplemental Data:

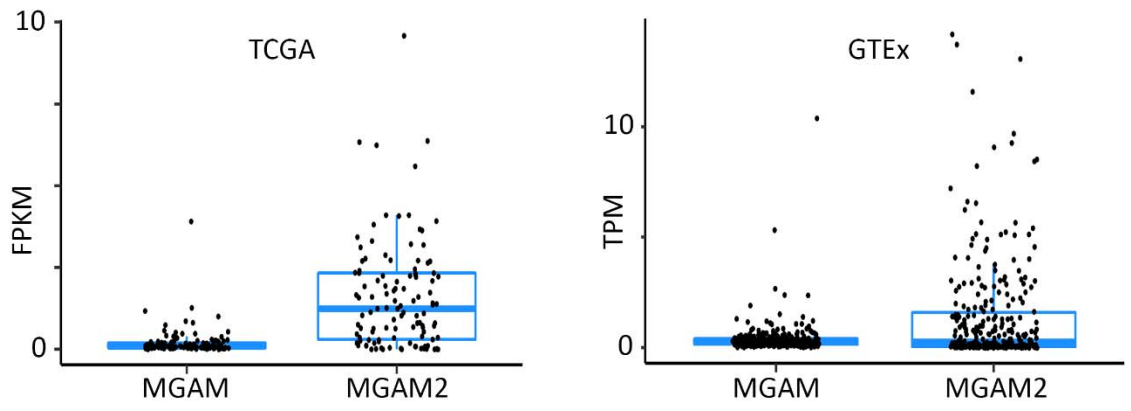

Figure S1-1. The distribution of *MGAM* and *MGAM2* expression in normal breast tissues samples from TCGA and GTEx.

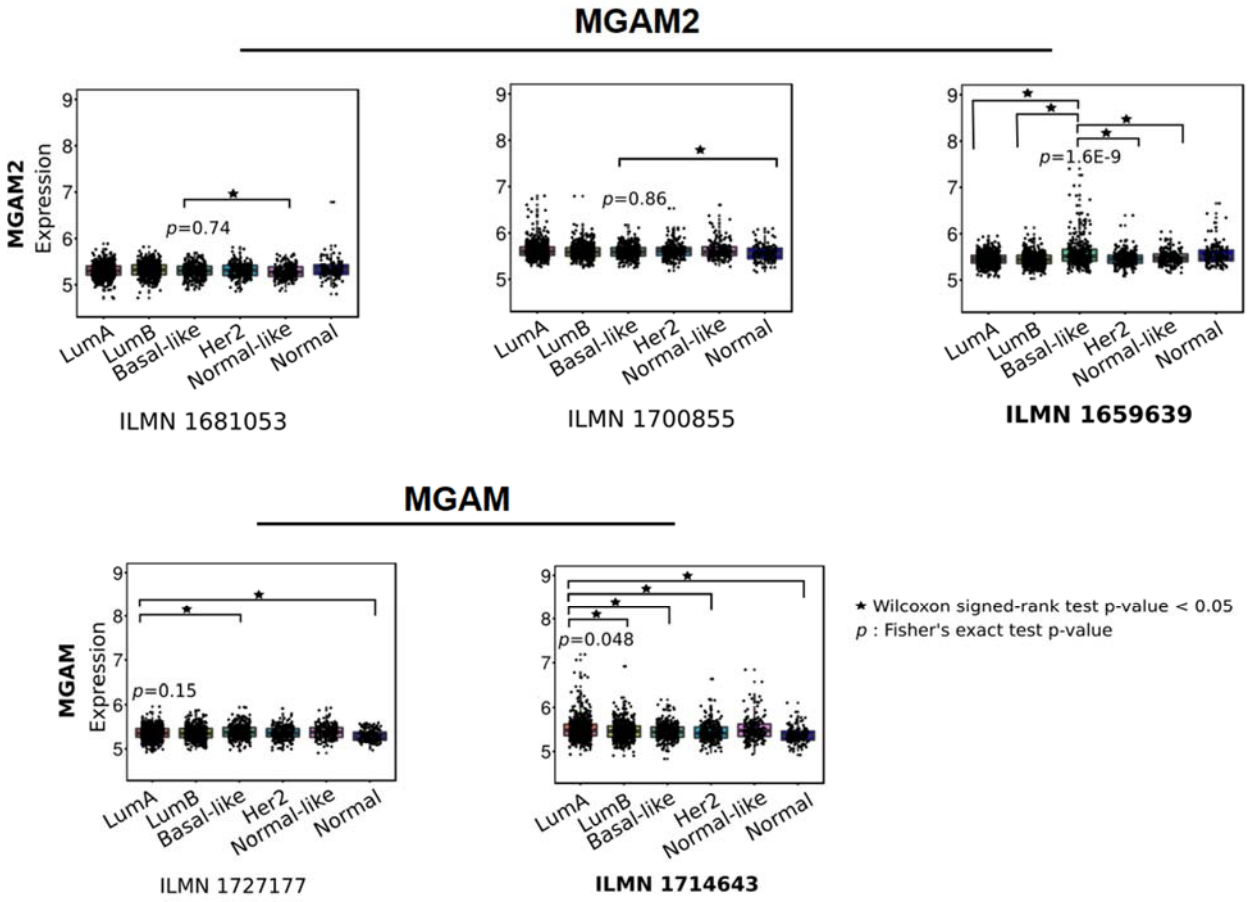

Figure S1-2. We examined the distribution of MGAM2 and MGAM expression among the 2000 breast cancers investigated by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). Unlike TCGA, METABRIC gene expression is determined with

microarray instead of RNA-seq, with MGAM2 and MGAM2 represented by three and two probes respectively. For MGAM2, one probe, locate at the 3'-end of the gene that are unique to MGAM2 (see later sections in the main text), shows significantly higher expression levels in BLBCs (p=1.6e-09) (Figure S1B and Table S1D). For MGAM, one probe also displays higher expression levels in LABCs (p= 0.05) (Figure S1B and Table S1D). These observations are consistent with TCGA study (Figure 1).

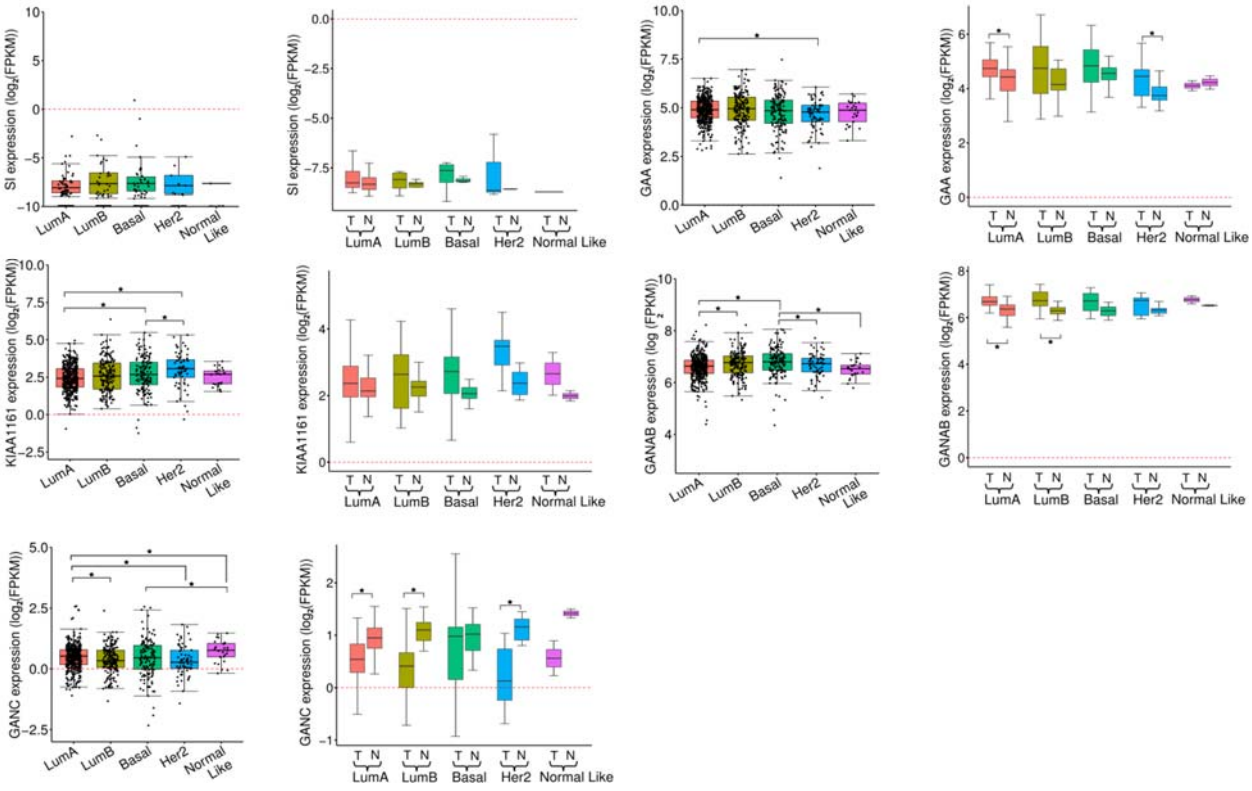

Figure S2A. Gene expression of other GH31 members in TCGA breast cancer subtype. See also Table S1A1.

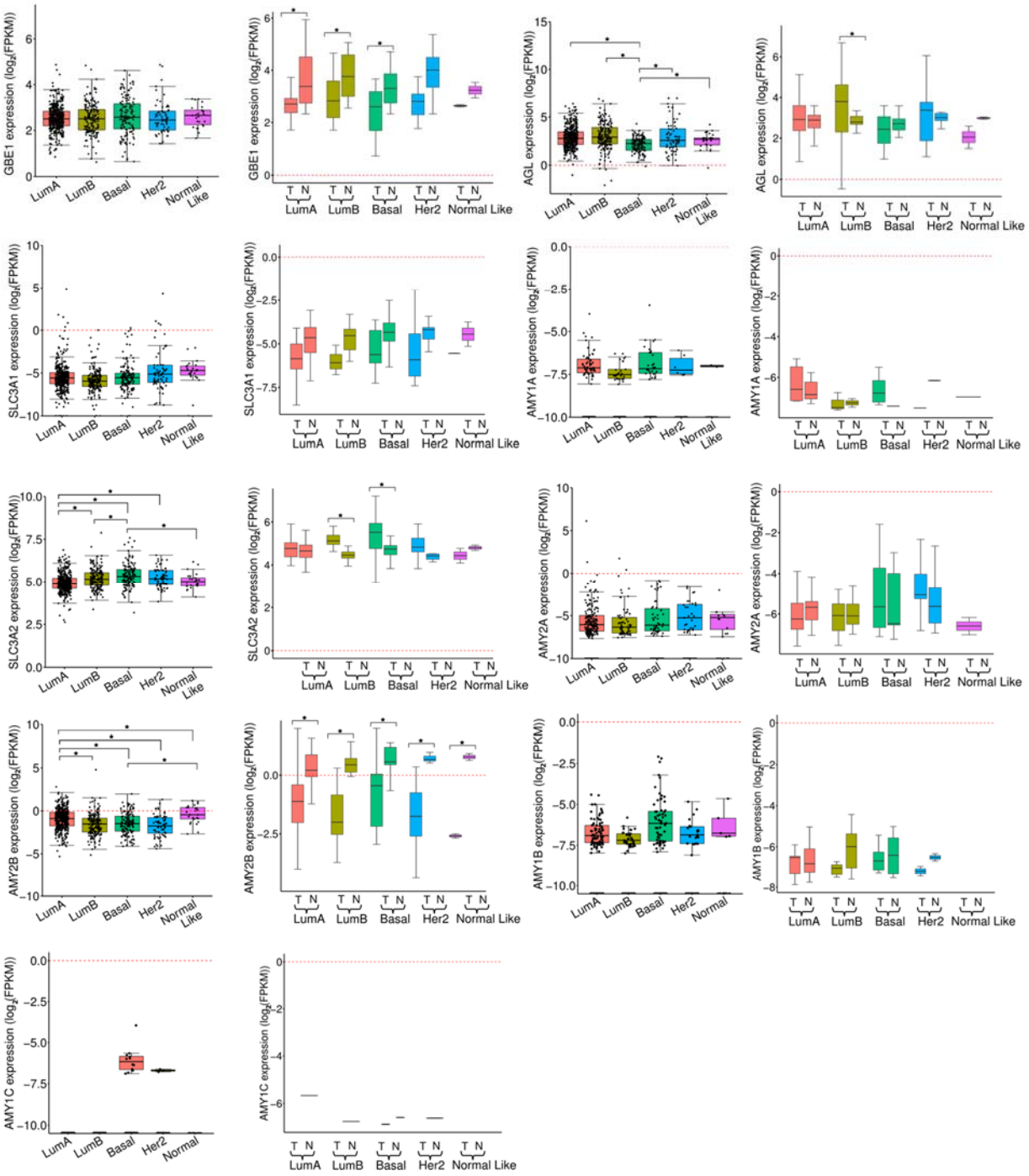

Figure S2B. Gene expression of GH13 members in TCGA breast cancer subtype. See also Table S1A1.

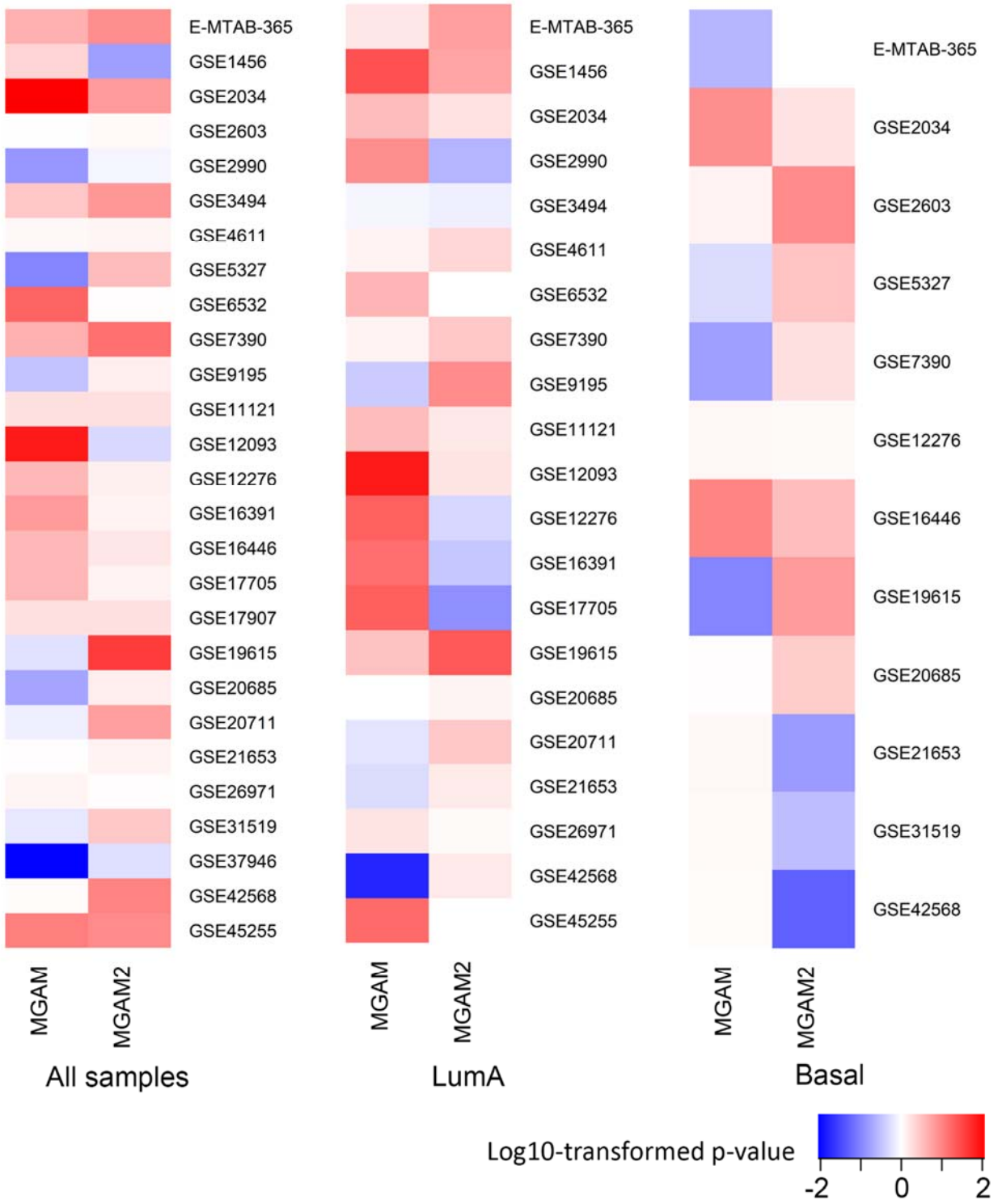

Figure S2C. Patient survival analysis with *MGAM* and *MGAM2* expression individual breast cancer datasets from the KM plotter site.

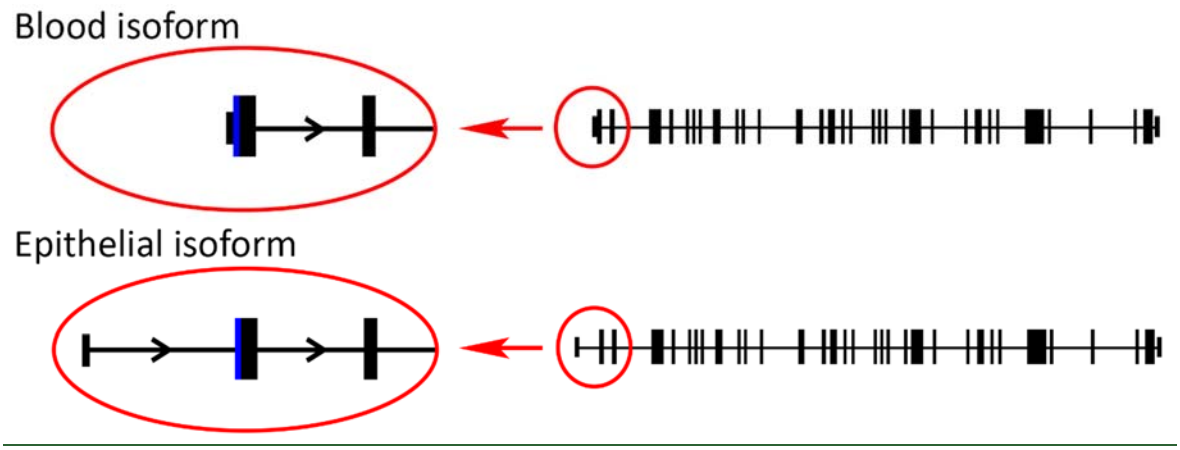

Figure S2D. We identified blood-specific and epithelial-specific alternative splicing (AS) forms of MGAM2, using RNA-seq data from GTEx. Blue color indicates the start codon. Taller bars represent coding exons while short bars represent UTR (untranslated region) exons, and lines between the bars indicate introns.

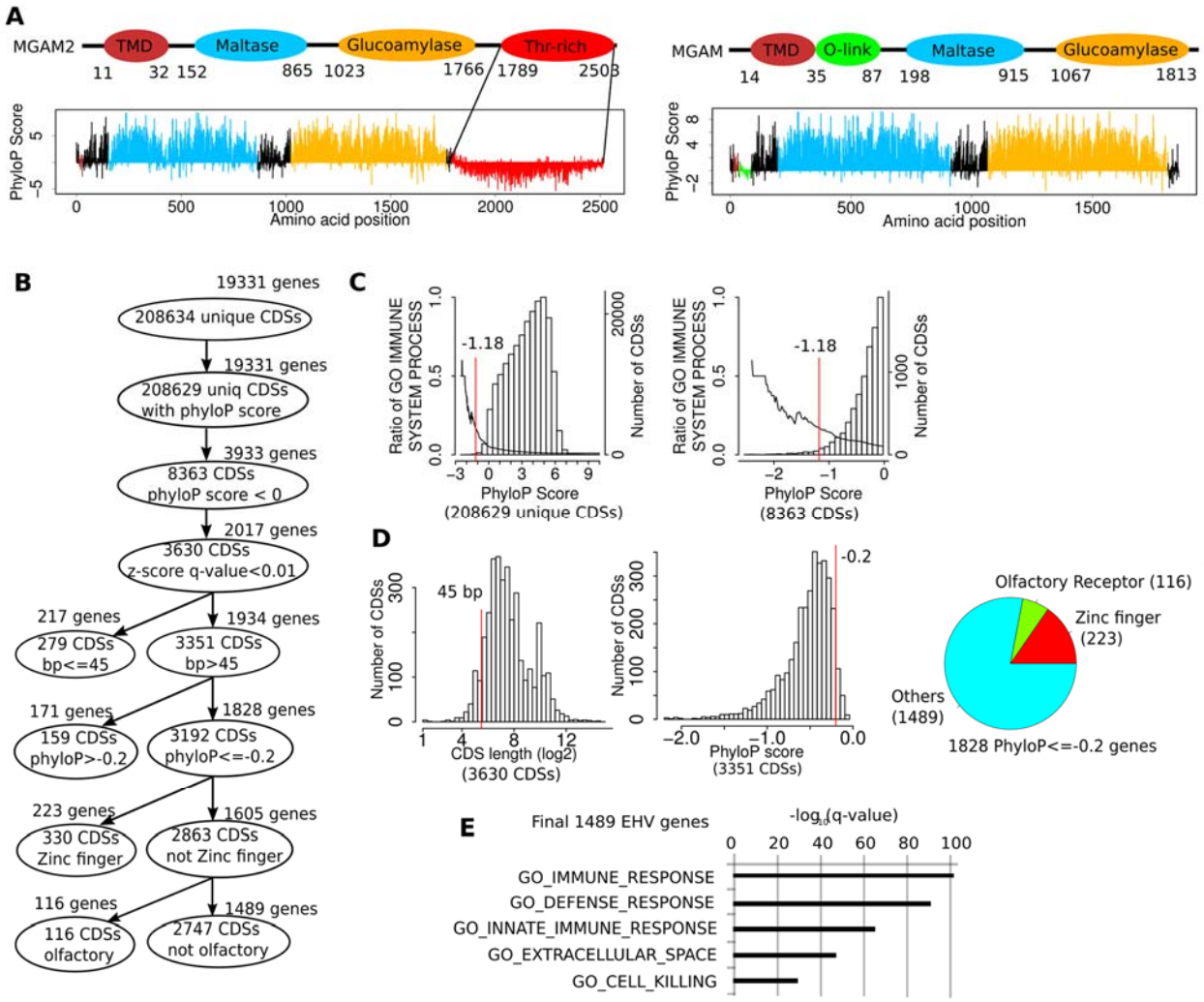

Figure S3. Analysis based on phyloP scores of 100 species. The figure is presented the same way as Figure 3.

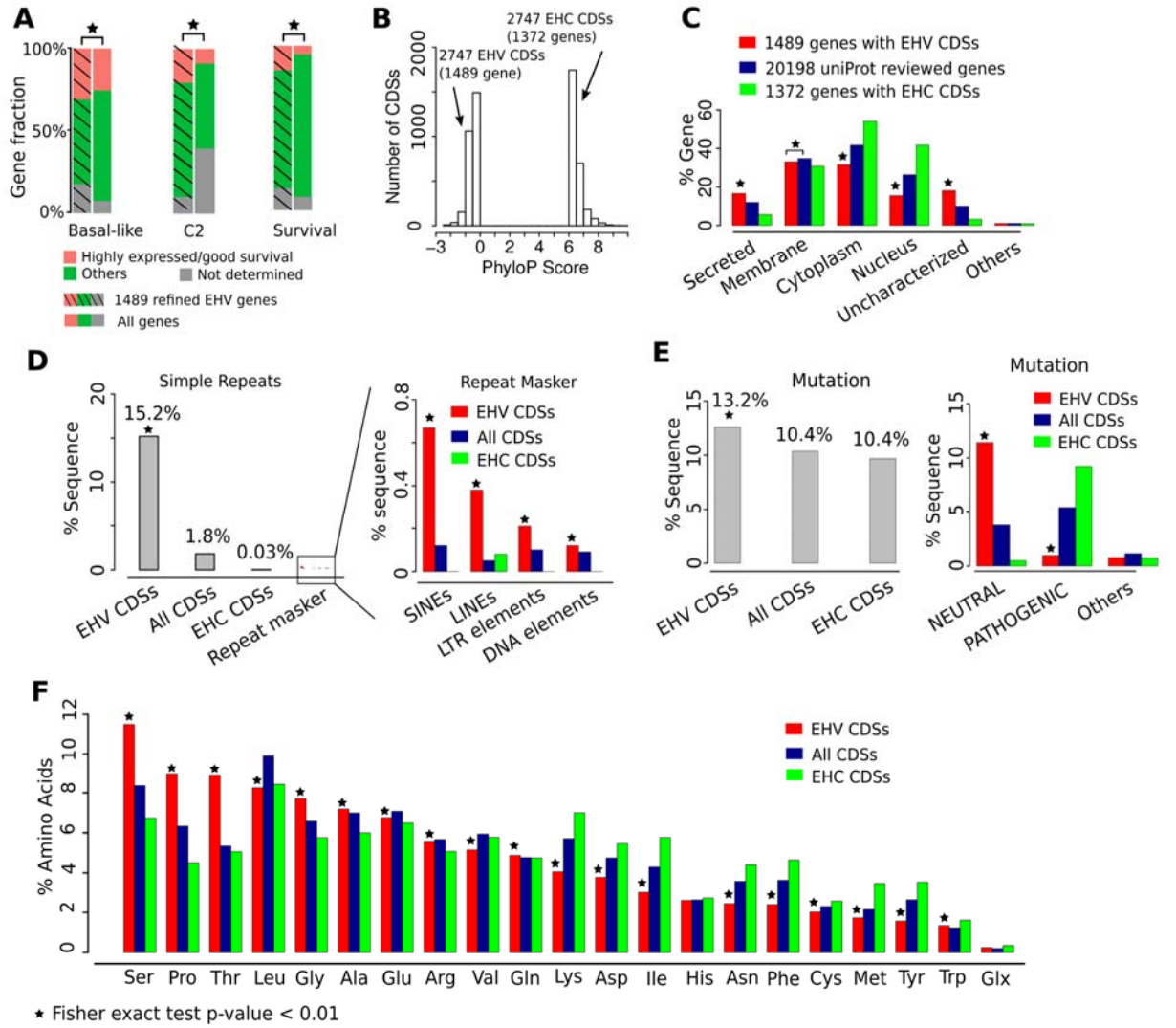

Figure S4. Analysis based on phyloP scores of 100 species. The figure is presented the same way as Figure 4.

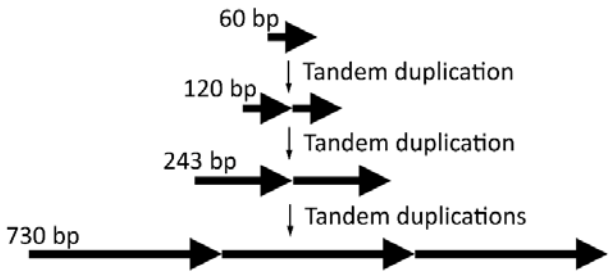

Figure S4-1. A total of 730bp sequence, encoding about 1/3 of the EHV domain of MGAM2, has arisen via several tandem duplications of a 60bp sequence, which has a consensus sequence of "CACTAATGCTACTGTTCCTATACAACCACACCTTTCCCAACAAGTACTACTAGTGCTA".

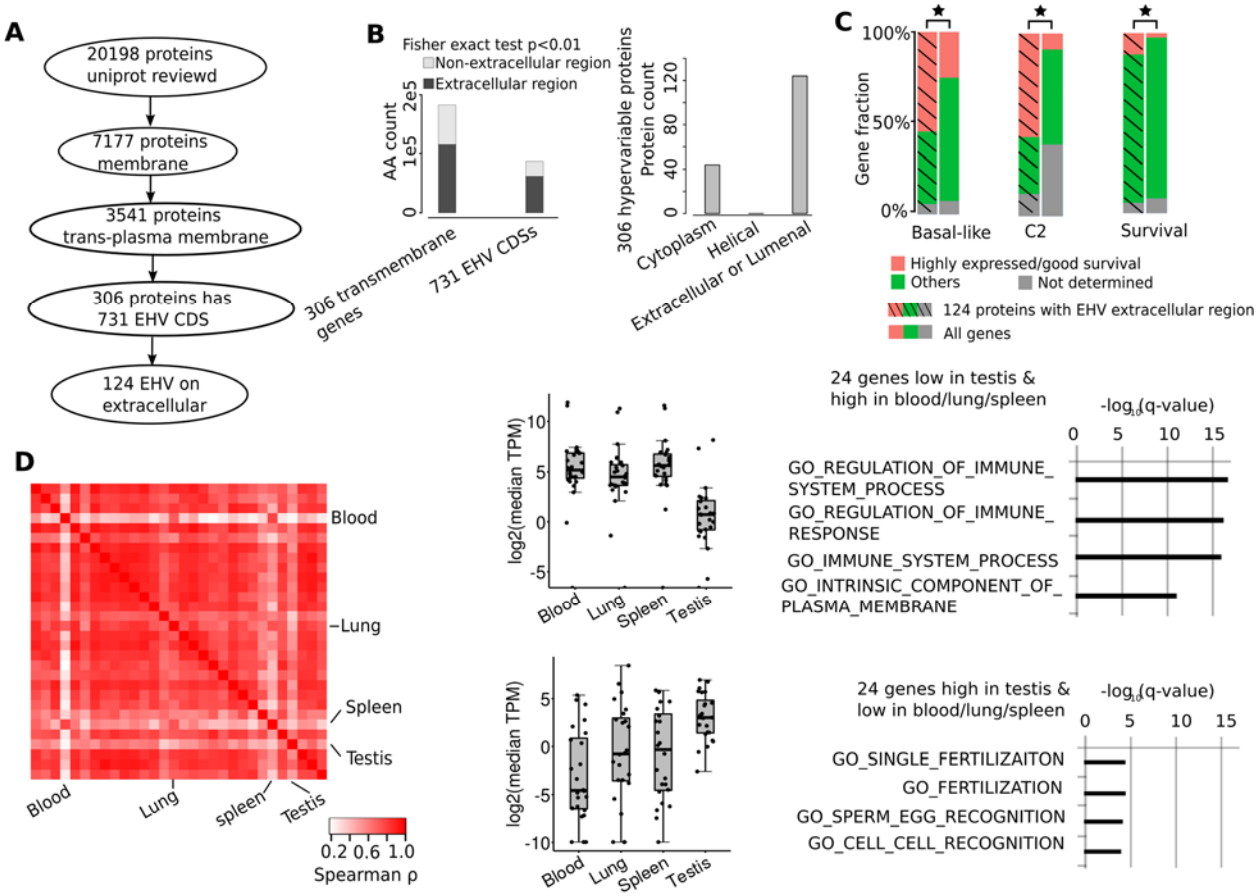

Figure S5. Analysis based on phyloP scores of 100 species. The figure is presented the same way as Figure 5.

Figure S6A. Sequence alignment between MGAM2, MGAM, SI, GAA and AMY1A at the active site. Red letters indicate key residues.

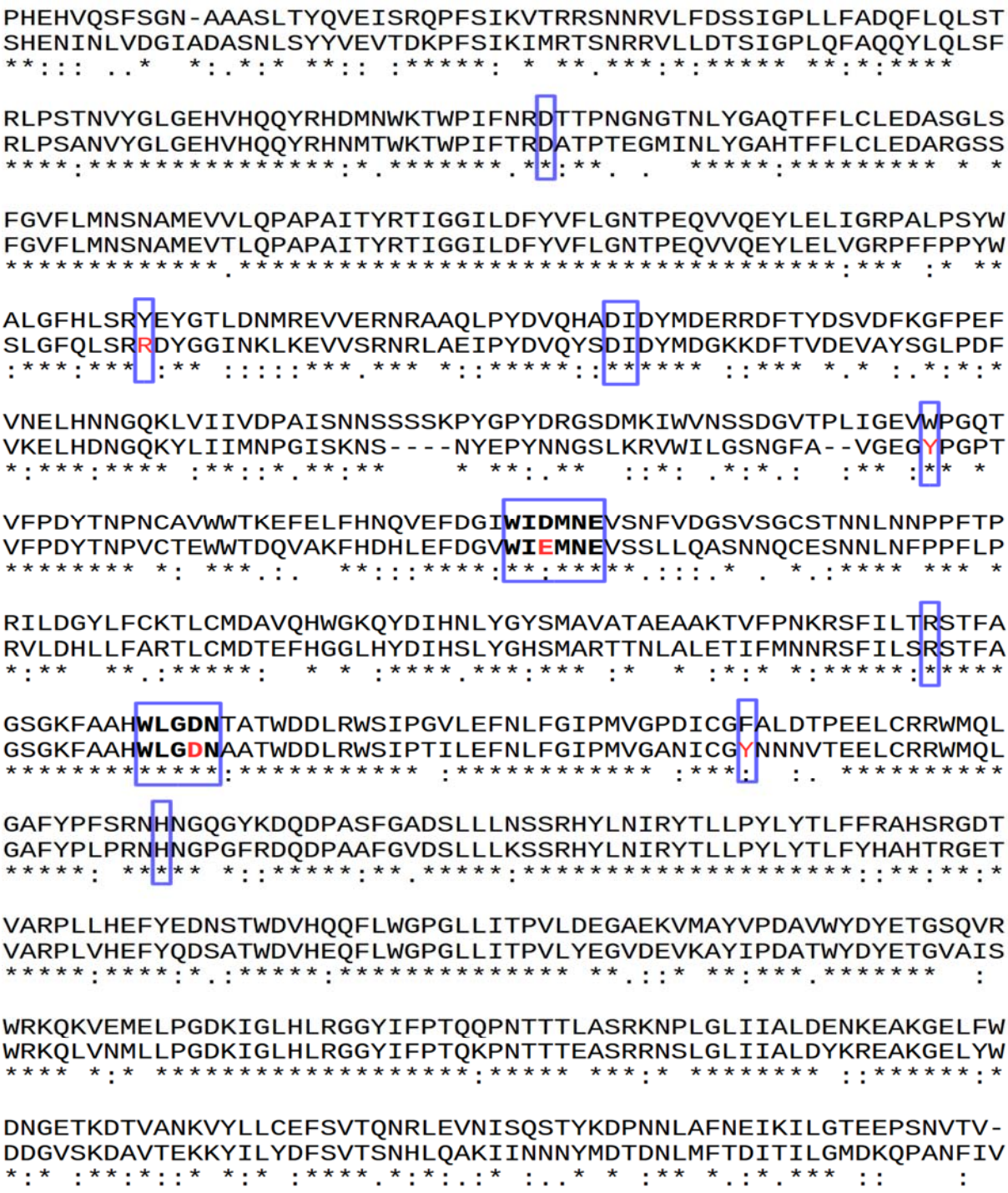

Figure S6B-1. Sequence alignment between MGAM (upper) and MGAM2 (lower) for maltase subunit. Rectangles indicate key amino acids at the substrate-binding site. Bold characters indicate signature peptides of GH31.

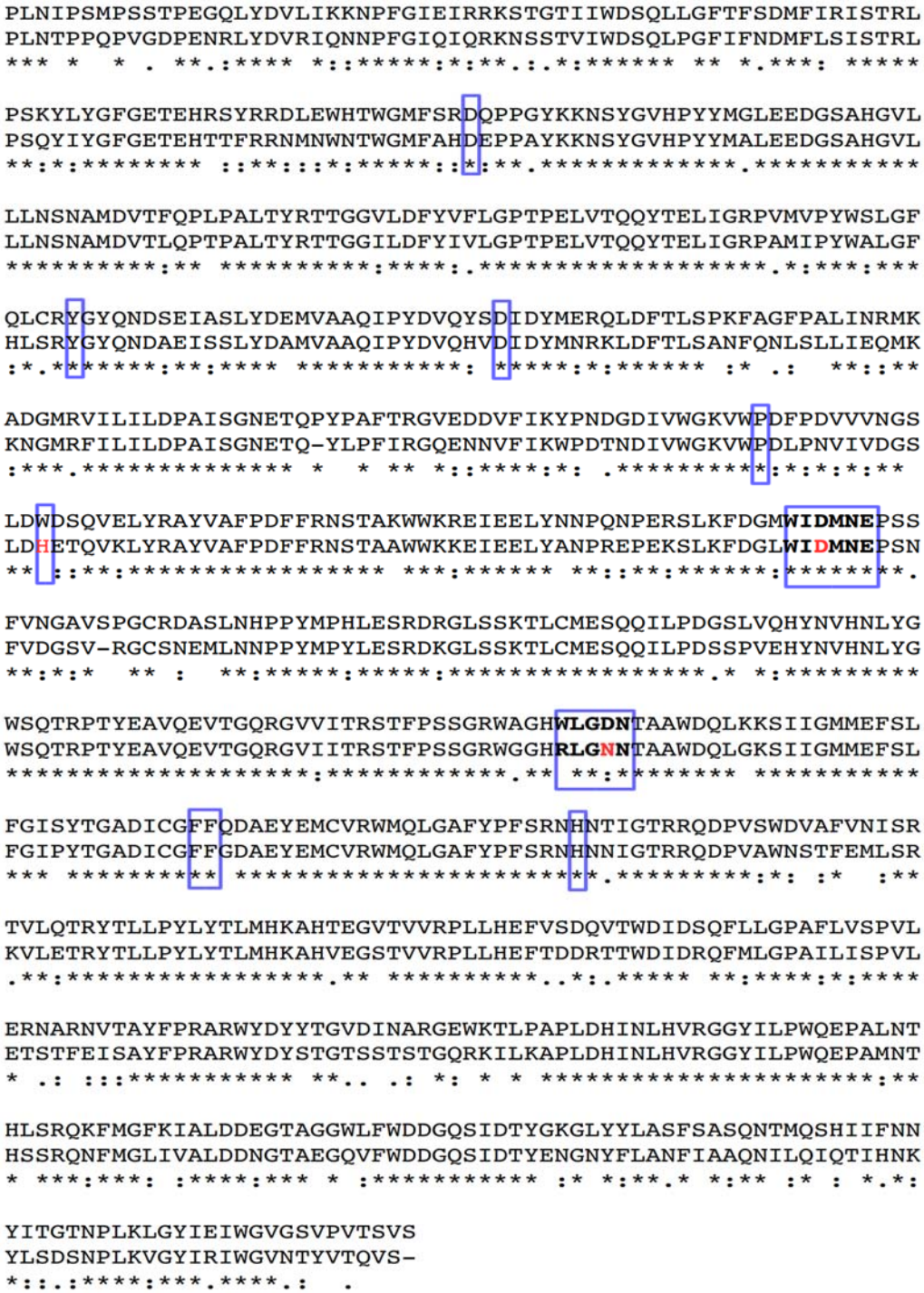

Figure S6B-2. Sequence alignment between MGAM (upper) and MGAM2 (lower) for glucoamylase subunit. Rectangles indicate key amino acids at the substrate-binding site. Bold characters indicate signature peptides of GH31.

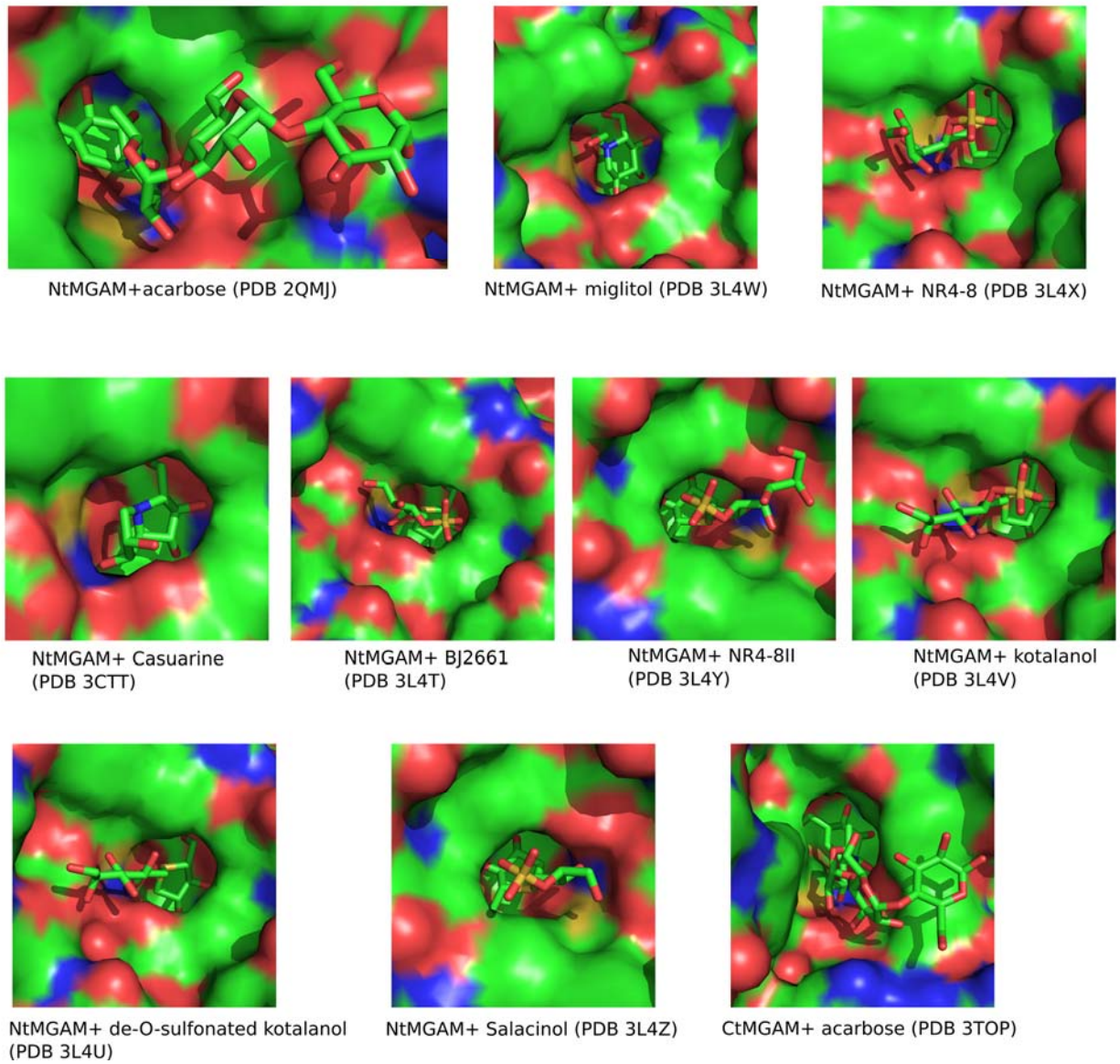

Figure S6C. Active site pockets of experimentally-determined structure of MGAM. Structures of MGAM in complex with small molecules are determined for N-terminal (2QMJ, 3L4W, 3L4X, 3CTT, 3L4T, 3L4Y,3L4V, 3L4U, 3L4Z) and C-terminal (3TOP) domains separately. The active site of MGAM is displayed in PyMol surface cavity mode, and the small molecules are shown in stick mode.

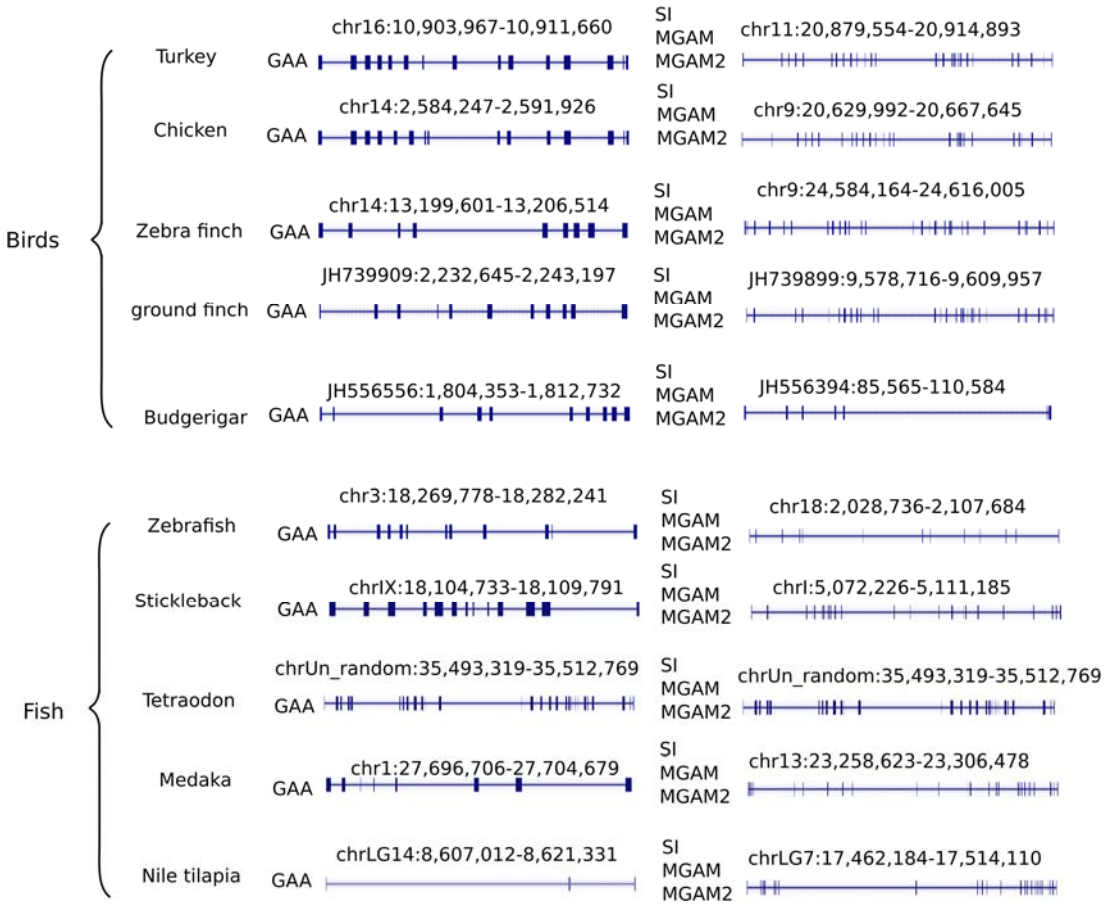

Figure S7. Gene structures of GAA and the ancestor of SI, MGAM, MGAM2 in fish and birds obtained from UCSC genome browser.

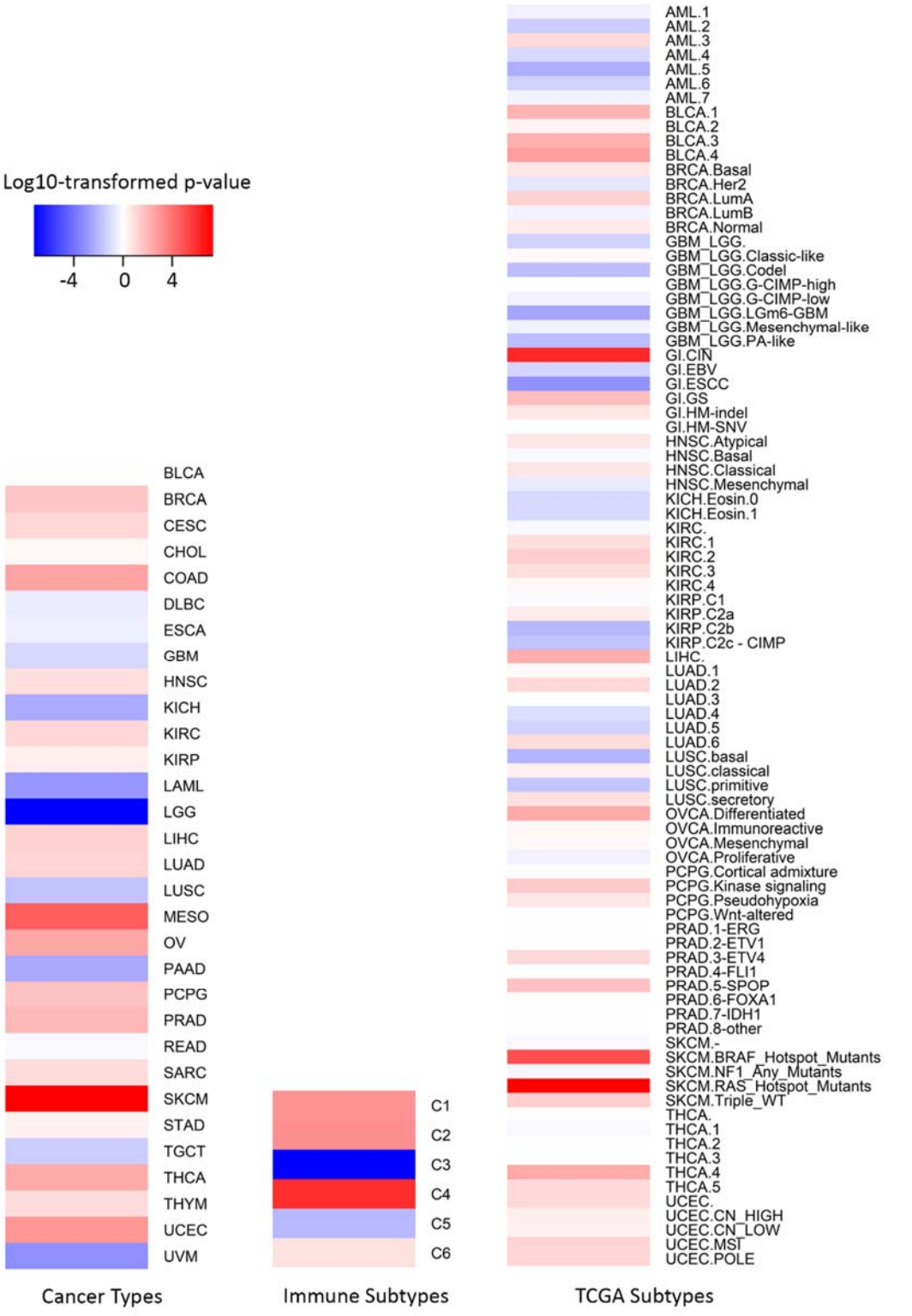

Figure S8a. Patient survival p-values. In order to study the >1000 EHV genes (Figure 3C) as a whole, we treated all EHV genes as a set and performed ssGSEA with EHV gene expressions in each TCGA cancer sample. We then conducted patient survival analyses with the ssGSEA scores. The results vary across cancer types. EHV genes are associated with better survival in

skin cancer (SKCM), *mesothelioma* (MESO) and colon cancer (COAD), but with worse survival in chromophobe renal cell carcinoma (KICH), acute myeloid leukemia (LAML), lower grade glioma (LGG) and uveal melanoma (UVM) ($p \leq 0.05$). Among cancer immune subtypes, EHV genes are associated with better survival in the C4 (lymphocyte depleted) subtype, but with worse survival in the C3 (inflammatory) subtype ($p \leq 0.05$). The results are largely supported by individual gene analysis.

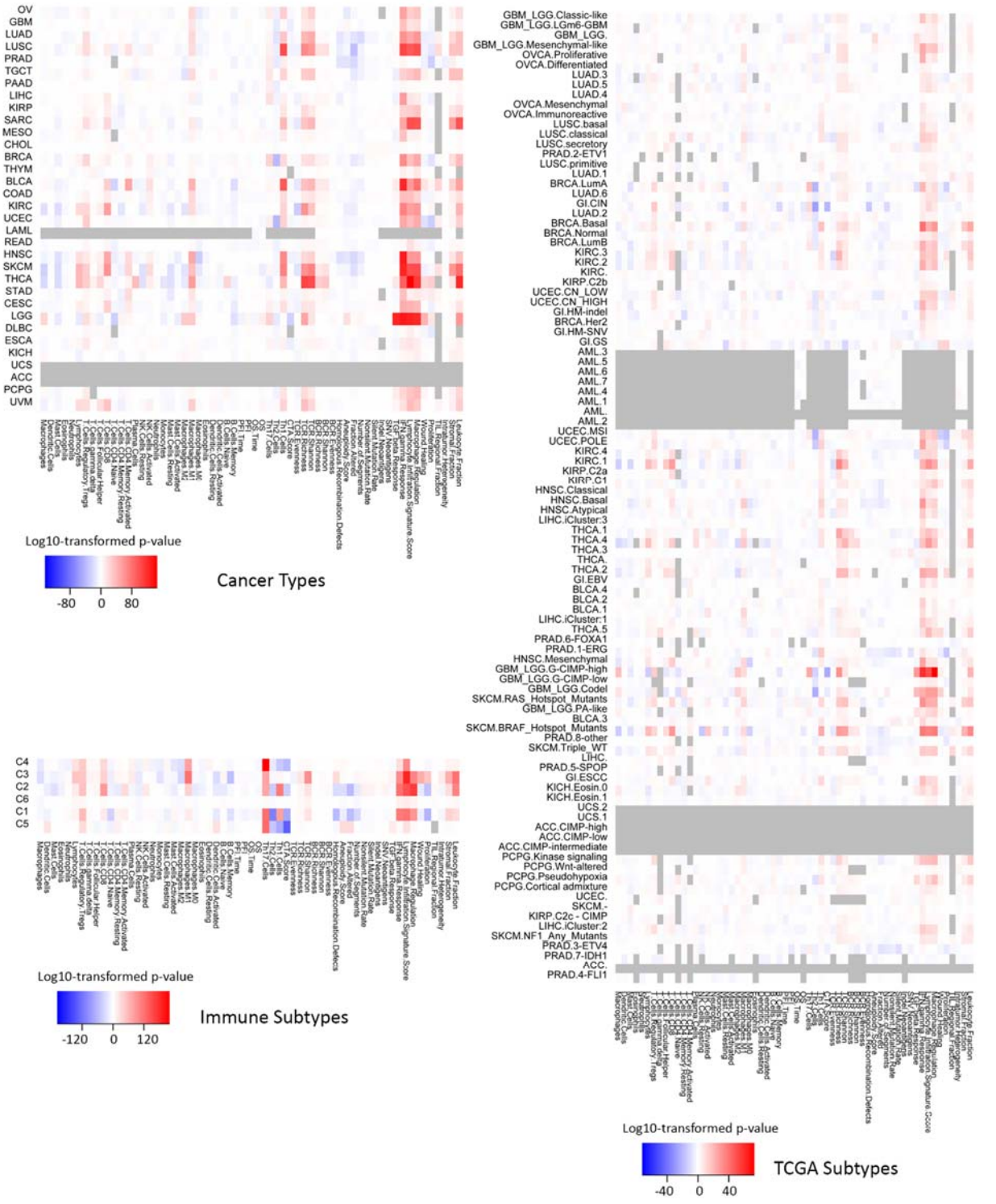

Figure S8b. Pearson correlation with ssGSEA scores. We also examined the association of ssGSEA scores with various immune features in TCGA samples. EHV genes are positively correlated with macrophage regulation, lymphocyte infiltration, IFN-γ response and TGF-β response across different cancer types. They are negatively correlated with cancer testis antigen (CTA) score and Th2 cells.

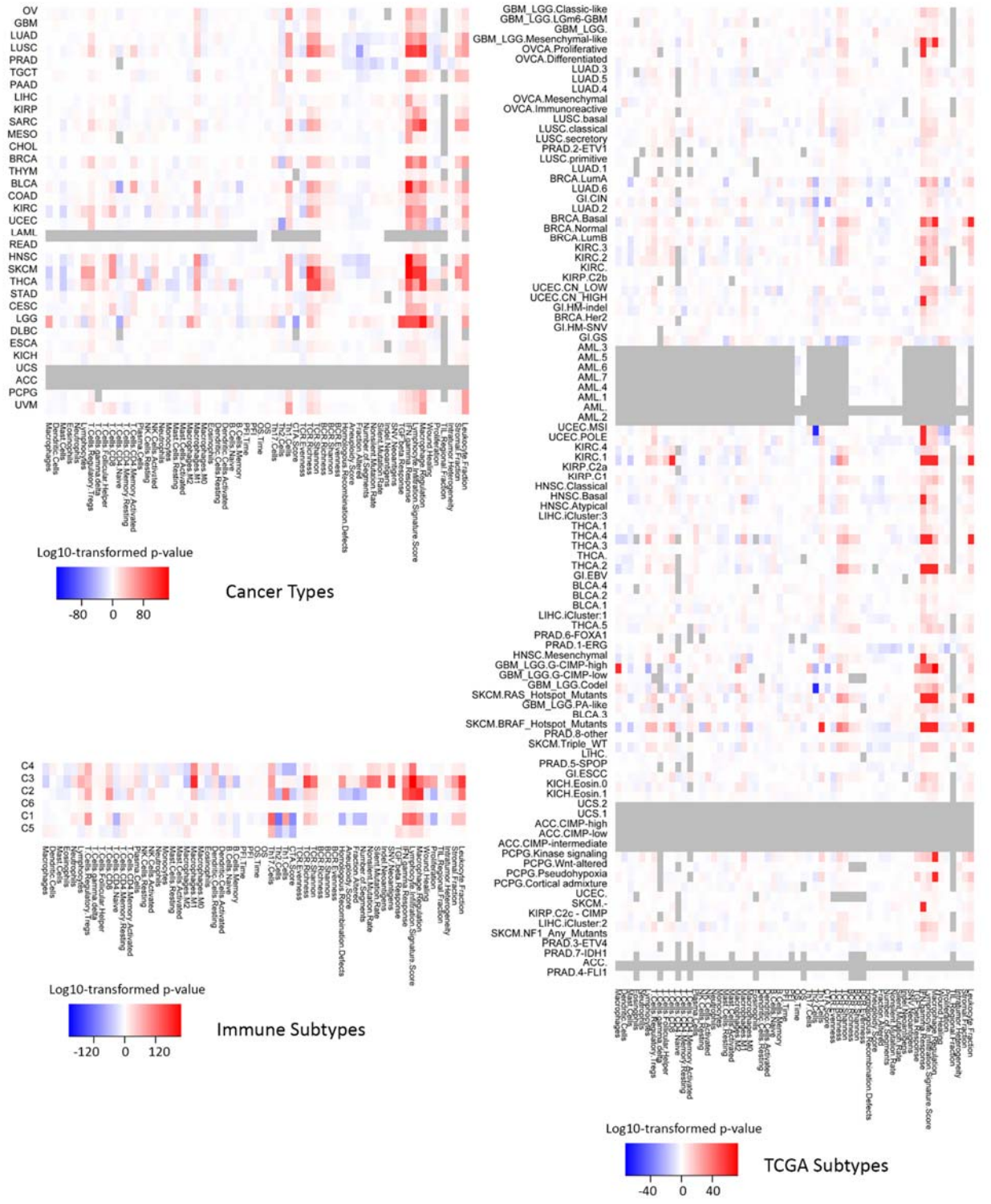

Figure S8c. Spearman correlation with ssGSEA scores.