

Protein Signatures for Distinguishing Colorectal Cancer Liver Metastases from Primary Liver Cancer Using Tissue Slide Proteomics

Xiaoman Zhou, Xiuyuan Wang, Ruizhen Bai, Hanjie Li, Dong Hua, Xiao-Dong Gao, Ganglong Yang and Quan Liu

Supplementary Data Supplementary Figures

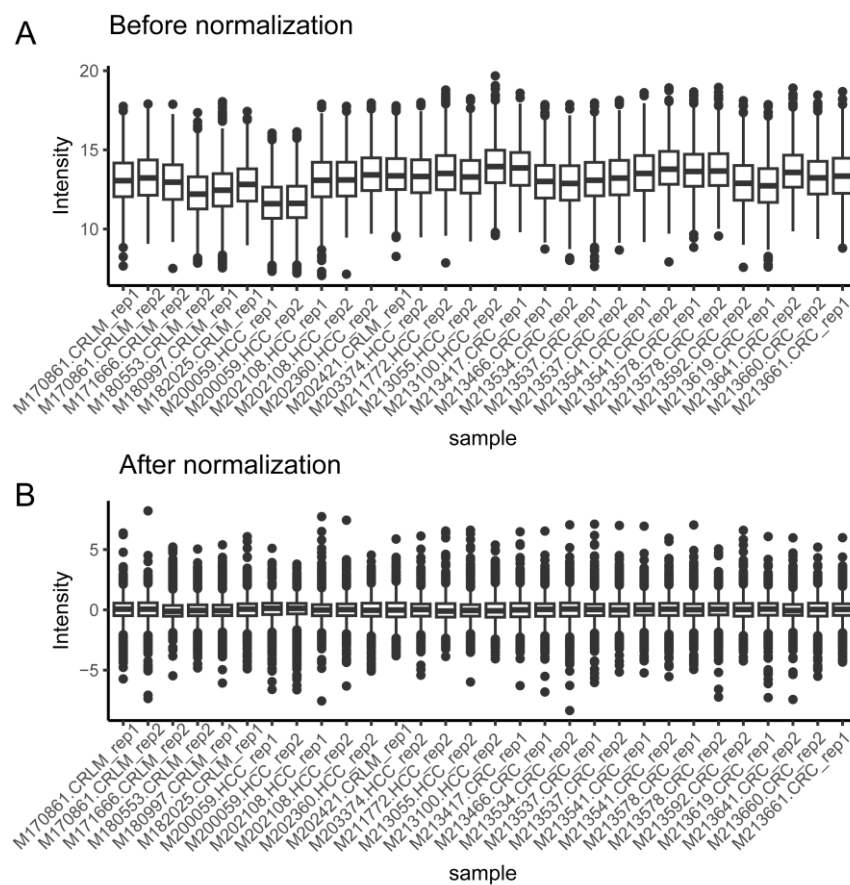
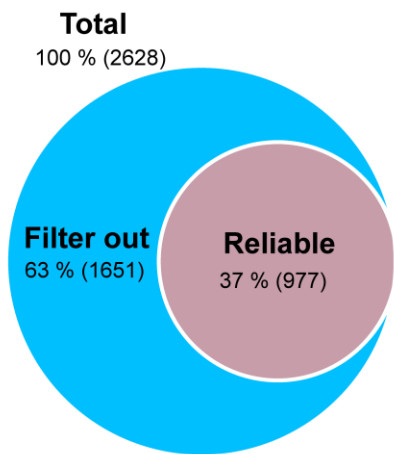


Figure S1. Protein intensity boxplot of tissue samples. Raw data for all samples were retrieved and quantified using MaxQuant software, followed by weighted merging by the reference channel (TMT_10). Intensity values were log2 transformed, and normalized across sample groups using z-scores normalization. Boxplots of 20 randomly selected samples before (A) and after (B) normalization were presented.

A



B

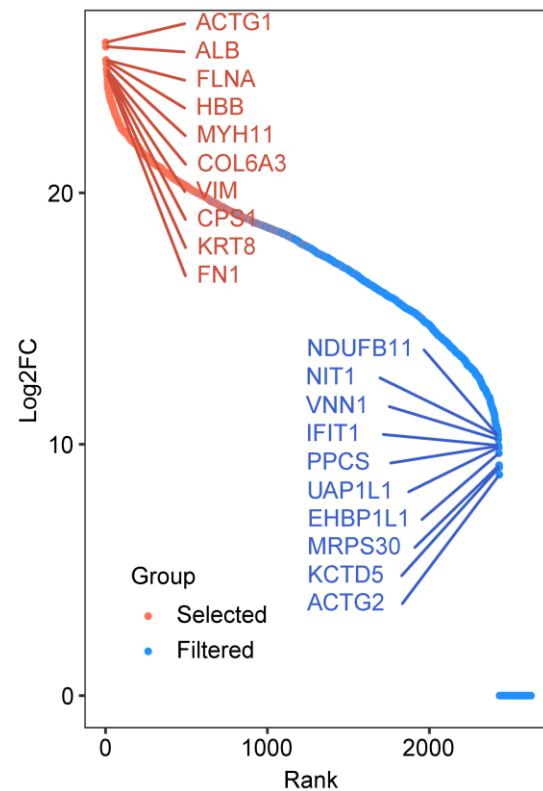


Figure S2. Reliable protein screening. (A) Wayne plot was generated by selecting proteins with score values (>10) and detection rates in each group of samples ($>70\%$). The chosen proteins accounted for 37% of the total identified ethical proteins. (B) Rank plot was constructed by sorting proteins according to their abundance, with red representing the reliable proteins selected and blue representing the filtered-out proteins. The top 10 and down 10 protein names were marked in the plot.

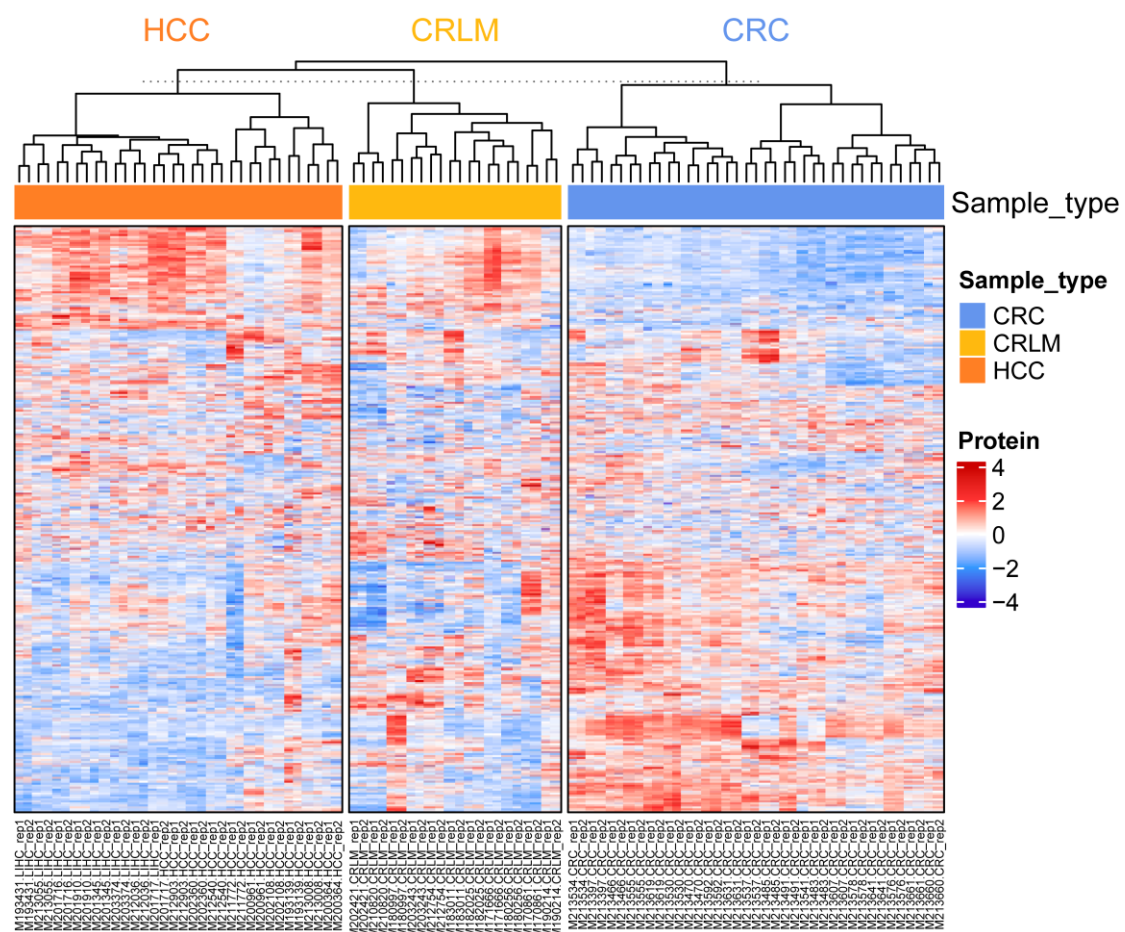


Figure S3. Heatmap of protein intensity in three sample groups. Columns: samples annotated by sample tissue type and arranged by cluster. Rows: Protein name arranged by cluster. Protein intensity was normalized by row z-score.

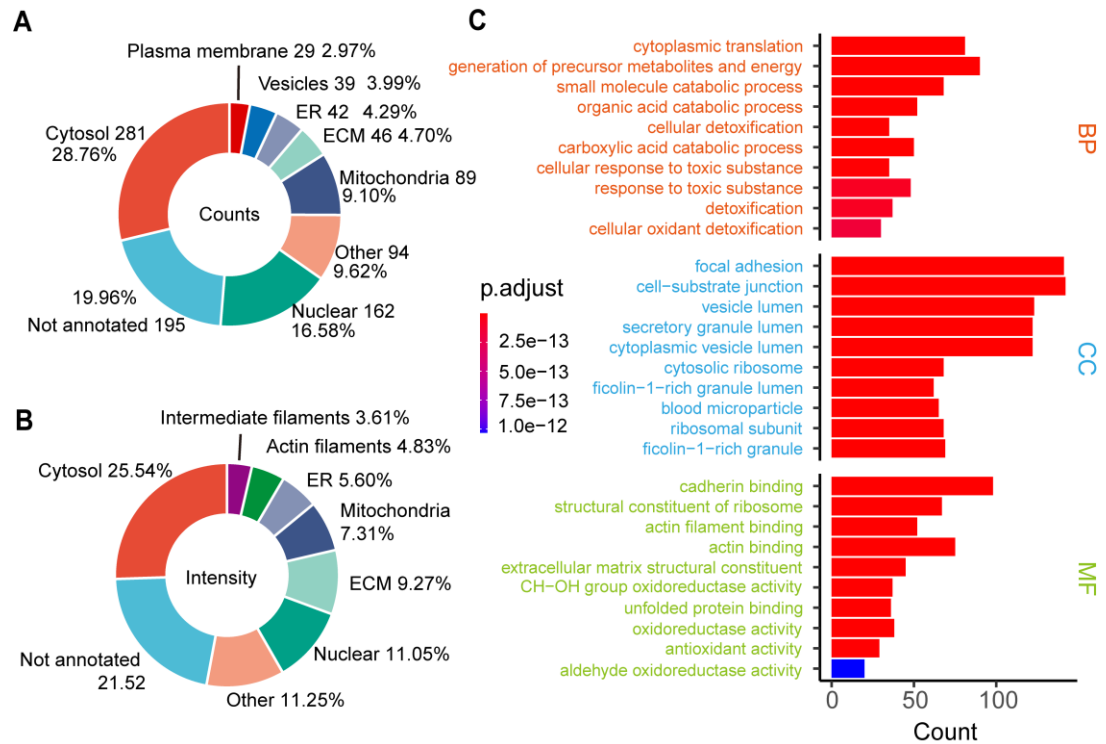


Figure S4. The location and functional analysis of detected proteins. (A) Pie chart of the count's proportion of identified proteins with subcellular localization. (B) Pie chart of the intensity proportion of identified proteins with subcellular localization. (C) Gene Ontology (GO) analysis of total proteins detected by LCMS. "BP" denotes Biological Process, "CC" denotes Cellular Component and "MF" denotes Molecular Function.

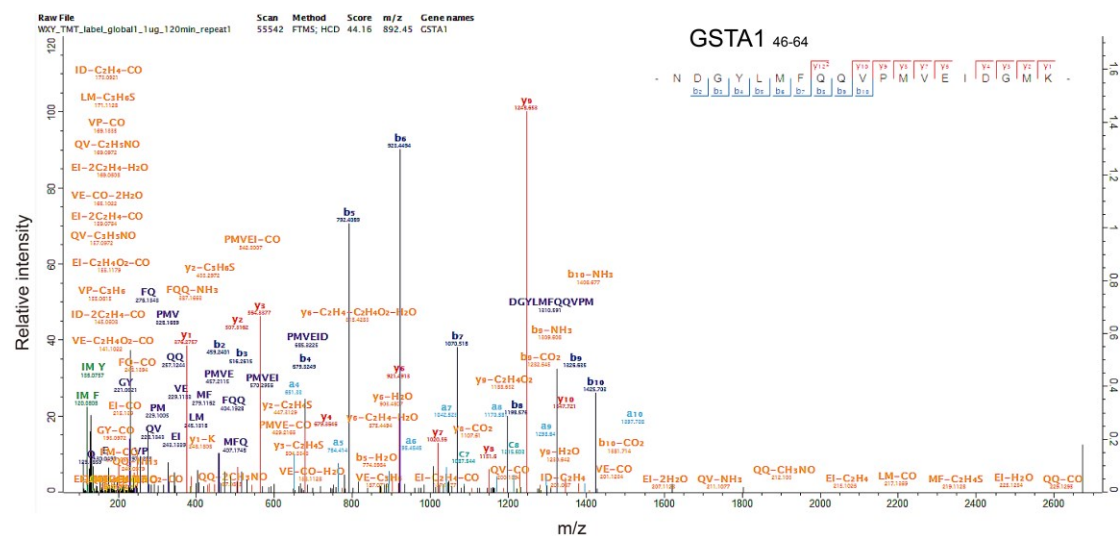


Figure S5. Identification and analysis of representative spectra for peptide segments by tandem mass spectrometry. Using a bottom-up approach, database retrieval and annotation were performed based on the peptide fragment mass-to-charge ratios. Uniquely identified protein-specific peptide segments were utilized to confirm the protein's identity. The figure above illustrates a representative spectrum from the identification results. Each peak in the spectrum has been well annotated, matching the 46-64 peptide segment of the GSTA1 protein. This segment serves as a characteristic peptide for GSTA1 and can be employed for the confirmation of proteins within this family.

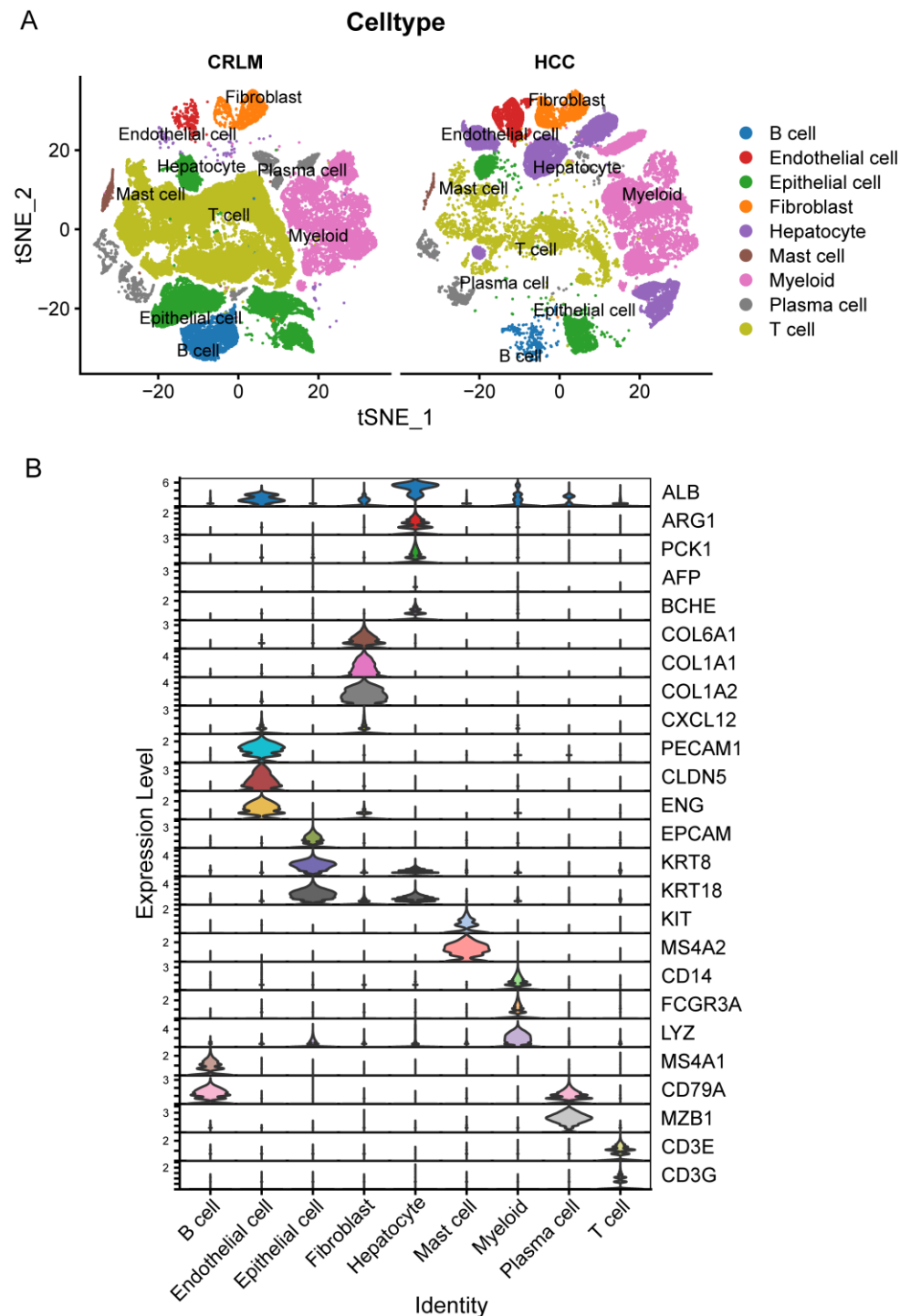


Figure S6. Single-Cell transcriptomes of CRLM and HCC. The marker genes selected were analyzed in single-cell transcriptomes of CRLM (n=6) and HCC (n=10), comprising a total of 90,149 cells. The raw data was obtained from GSE178318 and GSE149614, and processed using Seurat for integration and harmony for batch effect removal. Cell subtypes were annotated based on protein markers from the original paper. (A) The t-SNE distribution plot of single-cell groups. (B) The violin plot of cell type markers expression in each cell group.

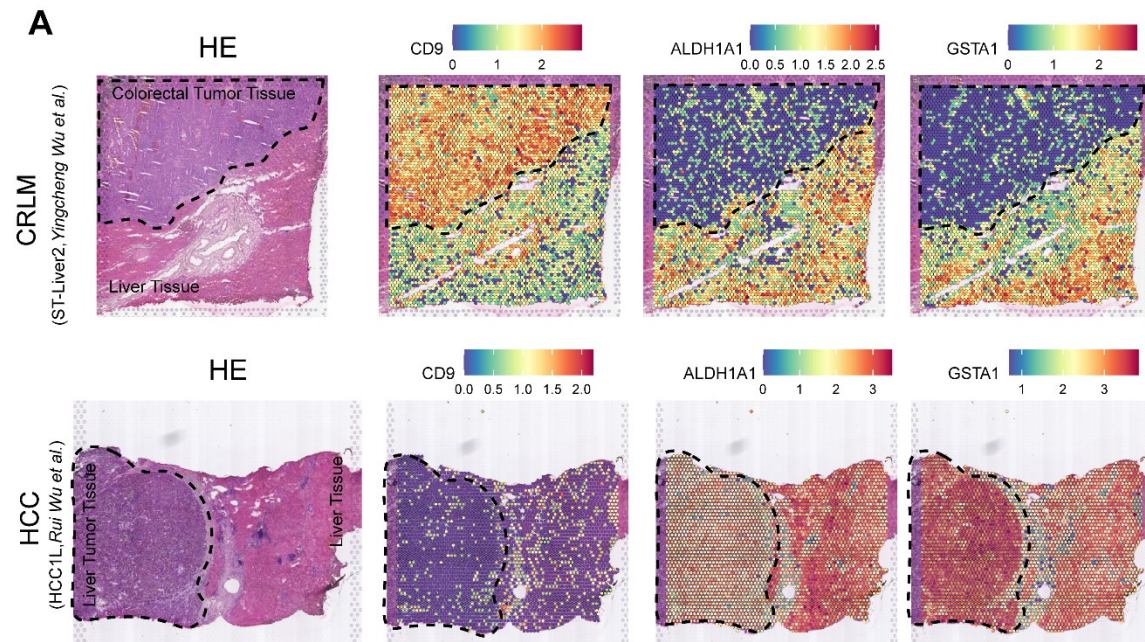


Figure S7. Validation of proteins obtained from the ML model in a spatial transcriptome dataset. The expression of CD9, ALDH1A1, and GSTA1 in tumor area (in dash circle) of CRLM and HCC. Raw data were from OEP001756 (ST-Liver2) and HRA000437 (HCC1L), and processed by Seurat.

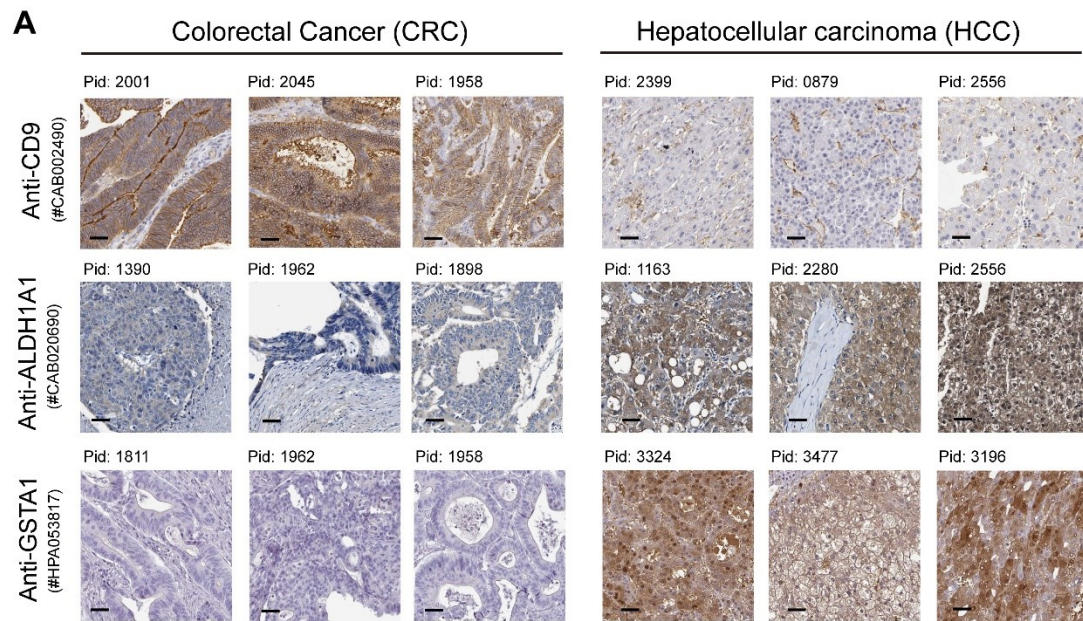


Figure S8. IHC analysis of marker proteins in CRC and HCC datasets. Images were from Human Protein Atlas with permission.