Cell Reports Medicine, Volume *4*

Supplemental information

Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood

Yuhuan Tao, Shaozhen Xing, Shuai Zuo, Pengfei Bao, Yunfan Jin, Yu Li, Mingyang Li, Yingchao Wu, Shanwen Chen, Xiaojuan Wang, Yumin Zhu, Ying Feng, Xiaohua Zhang, Xianbo Wang, Qiaoran Xi, Qian Lu, Pengyuan Wang, and Zhi John Lu

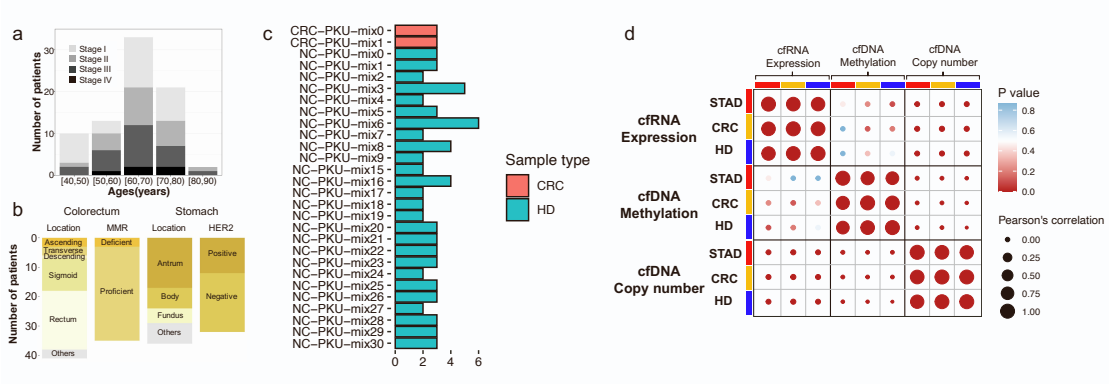

Figure S1. Clinical information and correlations of the multi-omics data, related to Figure 1.

a, Age and stage distribution of the patients. Age distribution of the patient cohort (range, 42–87 years; median, 64 years) highlighting disease stage.

b, Primary tumor location and molecular subtypes of the patients.

c, Illustration of the samples mixed by multiple individuals' plasma. Simultaneously sequencing cfDNA and cfRNA in the same individual consumes 2-3 ml plasma. In some cases, samples were mixed from individuals of same gender and similar age, due to the relatively small volume collected. More detailed usages of mixed samples are described in Supplementary Table 1.

d, Correlations of 77 samples with paired 3-omics data (cfWGS, cfMeDIP-seq, total cfRNA-seq).

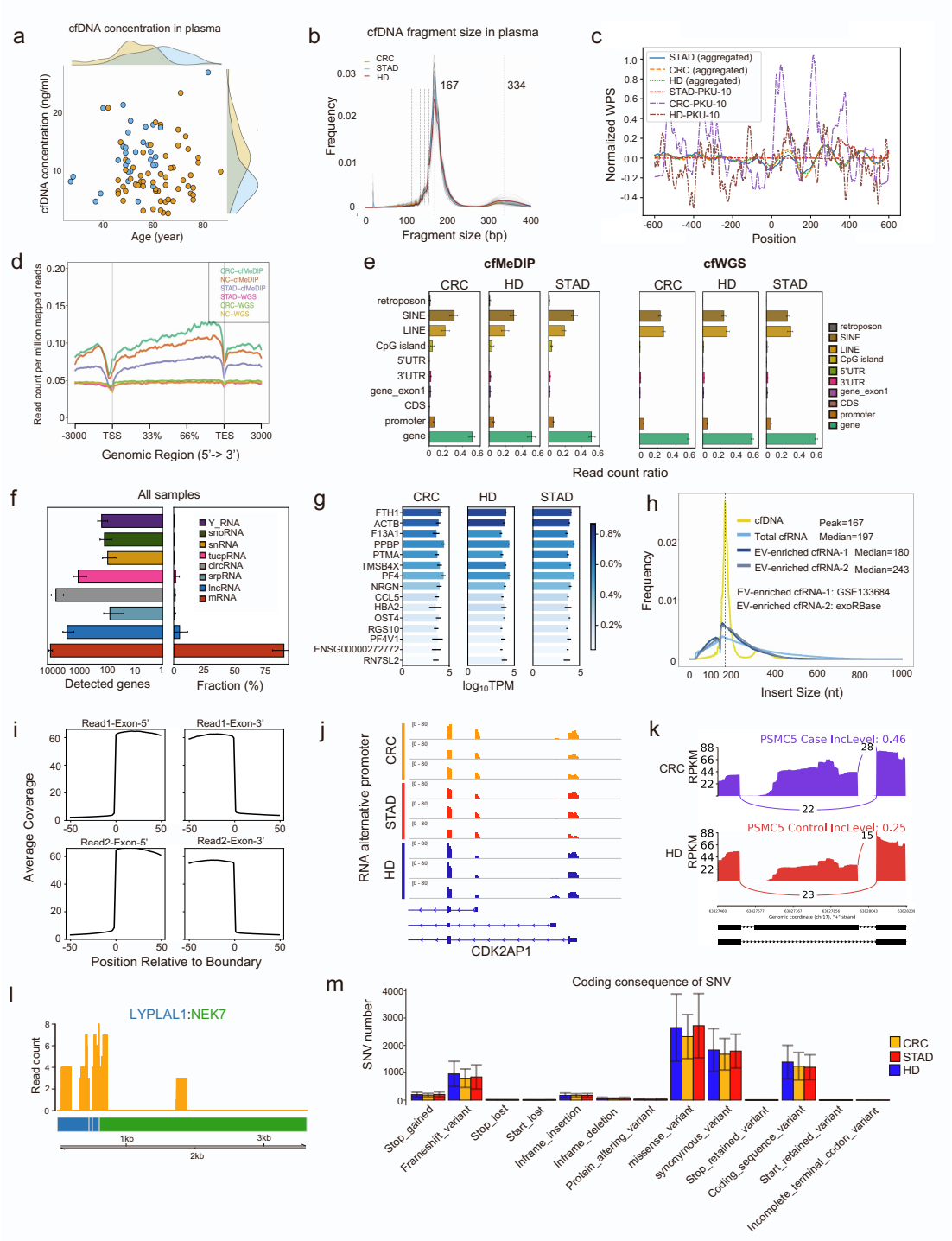

Figure S2. Basic characteristics of the cell-free multi-omics data, related to Figure 2.

a, Concentration of the cfDNAs (ng/ml) in plasma for individuals with different age. Orange: tumor group, blue: normal group, side graphs show global distribution.

b, Distribution of the cfDNAs' fragment sizes in plasma. Average cfDNA fragment size in 3 different groups: colorectum cancer patients (yellow), stomach cancer patients (blude) and health donors (red). The most frequent fragment size is 167 bp and a second prevalent fragment size is 334bp. The fragment size is highly consistent with nucleosome protected DNA length.

c, Distribution of the cfDNA derived windowed protection scores (WPS) around TSS.

d, Distribution of the cfWGS and cfMeDIP reads around TSS and TES. Y axis is RPM, x axis shows relative position to TSS and TES. RPM: read count per million, TSS: transcription start site, TES: transcription end site.

e, Distribution of the cfWGS and cfMeDIP reads in different genomic regions. Gene_exon1: representative first exon, Promoter: 2 kb upstream TSS to 0.5 kb downstream TSS, CDS: coding sequence, UTR: un-transcriptional region.

f, Detected gene number and fraction of reads for the total cfRNA-seq data. All RNA biotypes are annotated by GENCODE V27, except for tucpRNA, which is annotated by MiTranscriptome.

g, Top 15 genes (log_{10}TPM) sorted by the total cfRNA-seq reads. Color shows the fraction of total reads. Among the top expressed genes, we found housekeeping genes (ACTB/PTMA), platelet-related genes (F13A1/PPBP/PF4/PF4V1), erythrocyte-related genes (FTH1/HBA2) and exosomal RNAs, such as RN7SL2 (a signal recognition particle RNA).

h, Insert size for cfRNA, evRNA (extracellular vesicle) and cfDNA. The average insertion length of cfRNA is ~200 nt, which is different from cfDNA. Light blue: our total cfRNA-seq data. Dark blue: small evRNA (sEV-RNA-1) data downloaded from exoRBase version 1.0 [1]. Grey: small evRNA (sEV-RNA-2) data downloaded from GSE133684. Yellow: our cfWGS data.

i, Aggregating coverage of the total cfRNA-seq reads at the exon-intron boundaries. The sharp edge suggests limited DNA contamination in the total cfRNA-seq data.

j, Example: total cfRNA-seq reads showing the promoter usage of CDK2AP1.

k, Example: total cfRNA-seq reads showing a spliced exon of PSMC5.

l, Example: total cfRNA reads showing the fusion of LYPLAL1:NEK7.

m, Coding consequences of the cfRNA-derived single nucleotide variations (SNVs).

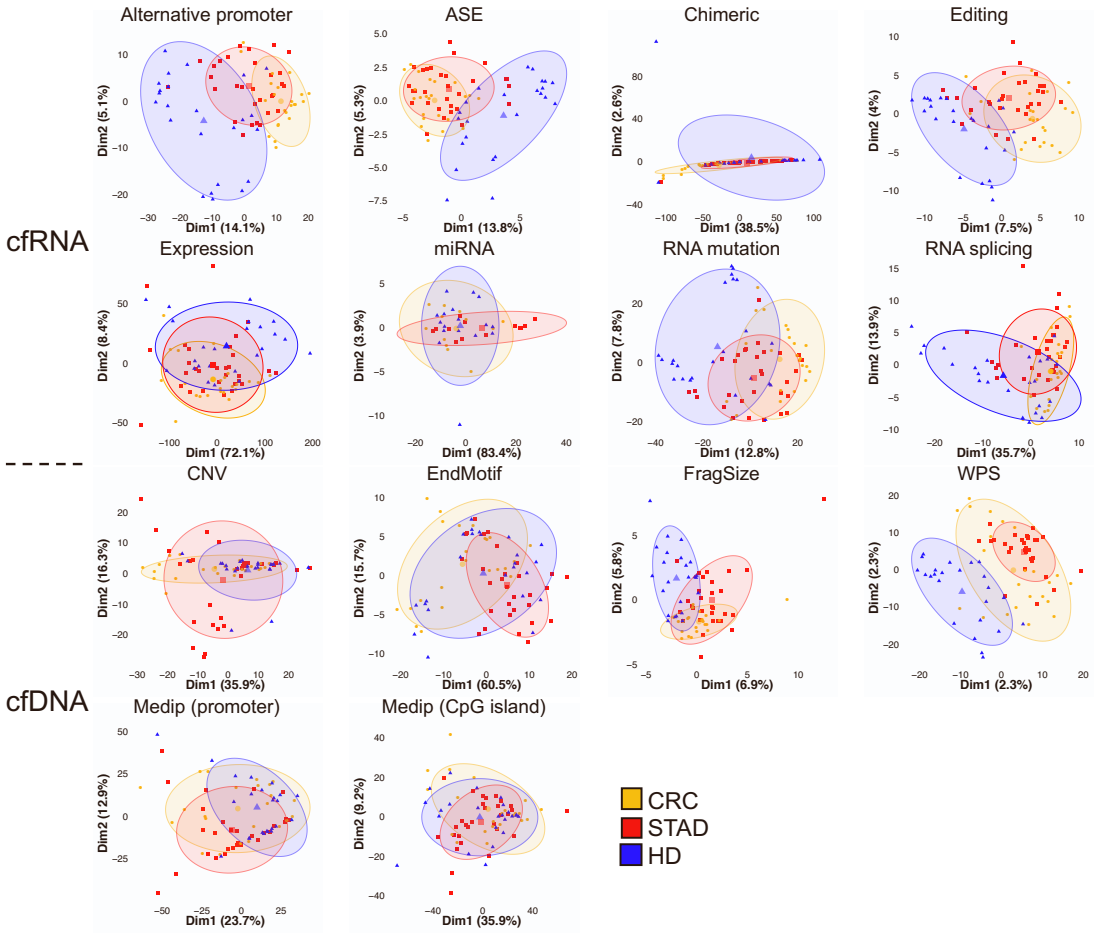

Figure S3. Principle component analysis of the differential alterations, related to Figure 3.

PCA analyses of all the differential alterations, including colorectal cancer (CRC) versus healthy donors (HDs), stomach cancer (STAD) versus HDs, and CRC versus STAD.

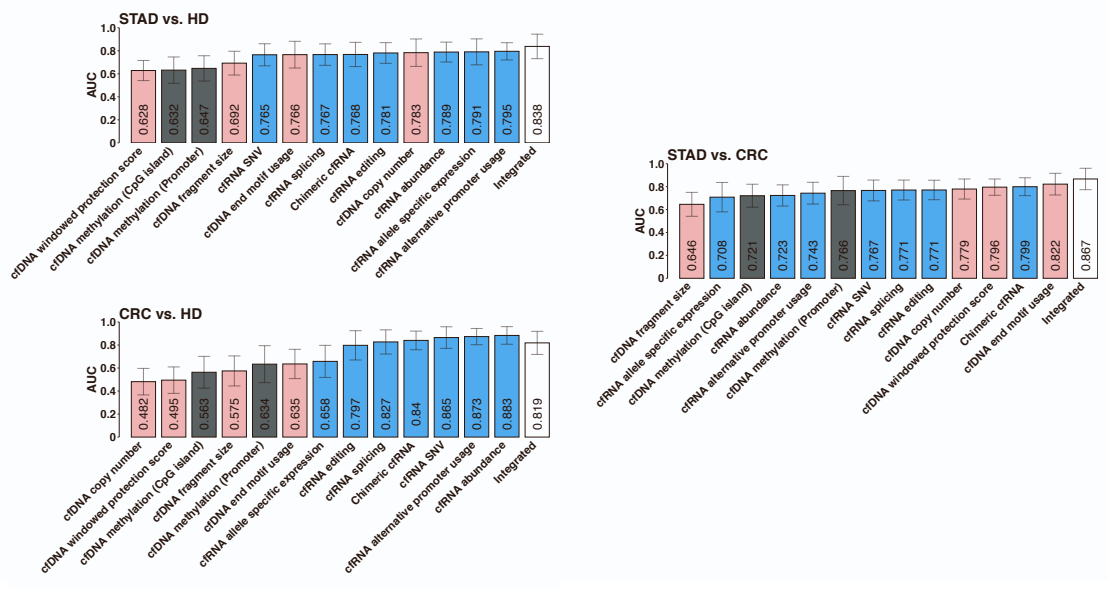

Figure S4. Classification performance of single-omics and multi-omics integrated variations at sample level, related to Figure 3.

The machine learning classification is based on random forest model in R package *randomForest* [2] and validated by bootstrap method. Each bootstrap, a dataset of 50 individuals is resampled as training set. For each single-omics variation, top 300 differential alterations in the training set are selected based on *P*-value. The integrated model is an ensemble of the probability of each single-omics model. The error bars represent the standard deviation of AUC. AUC: Area under curve. Usually, integrating more feature/omics would help the classification task, like STAD vs. HD, STAD vs. CRC here. In some cases, if some feature/omics were too noisy, it may hurt the overall performance, like CRC vs. HD here. However, the robustness of multi-omics' classification performance at sample level requires a large-size cohort (thousands of samples) to get a conclusive result.

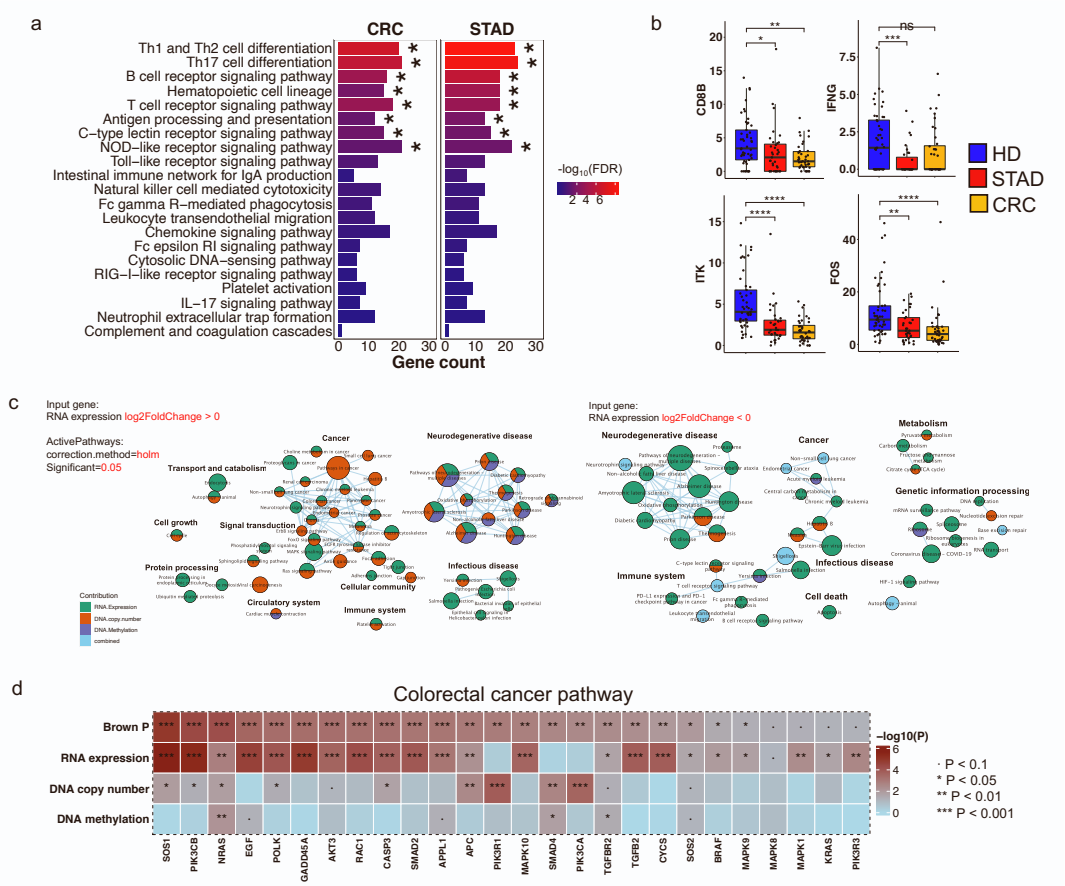

Figure S5. Downregulated immune pathways and genes defined by the total cfRNA-seq data, related to Figure 4.

a, Immune related pathways enriched in CRC and STAD patients' plasma using the total cfRNA-seq data. *False Discovery Rate (FDR) < 0.05.

b, Example genes downregulated in the cancer patients' plasma. *CD8B:* a cell surface glycoprotein found on most cytotoxic T lymphocytes. *IFNG:* its primary producers are effector T cells and NK cells. *ITK:* caused a cascade transcriptional effect in stimulated T cells and diminished FOS (AP-1 gene family) in cancer patients [3].

c, Integrated networks of the enriched pathways. The networks were derived by *ActivePathways* [4] based on the multi-omics data. All circles' Holm's-method-corrected *P*-values are less than 0.05. Circle size represents gene number in each pathway; edge represents the similarity between each circle larger than 0.375; *P*-value of each gene represents significance of differential alteration.

d, Example genes altered at different omics in colorectal cancer pathway. Brown P was an integrated *P*-value of cfDNA copy number, cfDNA methylation (promoter) and cfRNA abundance by Brown's method. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

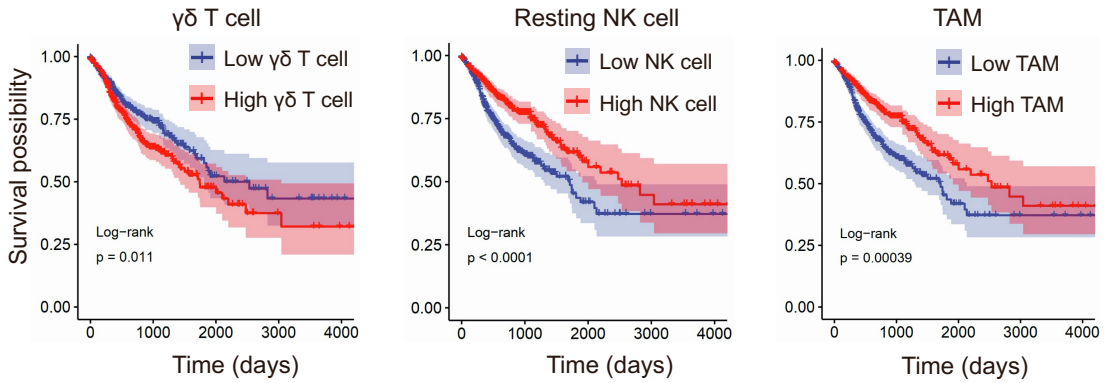

Figure S6. Survival analysis of cell type signatures significantly correlated with cancer stage in plasma, related to Figure 6.

γδ T cell, resting NK cell and TAM signatures are also significantly reveal the patients' survival time in TCGA cohort (1,006 colorectal and stomach cancer patients). NK cell: nature killer cell. TAM: tumor associated macrophage.

References

1. Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., Zheng, Q., Li, Y., Wang, P., He, X., and Huang, S. (2018). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. Nucleic Acids Research *46*, D106-D112. 10.1093/nar/gkx891.
2. Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32. 10.1023/A:1010933404324.
3. Gallagher, M.P., Conley, J.M., Vangala, P., Garber, M., Reboldi, A., and Berg, L.J. (2021). Hierarchy of signaling thresholds downstream of the T cell receptor and the Tec kinase ITK. Proceedings of the National Academy of Sciences *118*, e2025825118. 10.1073/pnas.2025825118.
4. Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N.S., Zhu, H., Abd-Rabbo, D., Mee, M.W., Boutros, P.C., Abascal, F., Amin, S.B., et al. (2020). Integrative pathway enrichment analysis of multivariate omics data. Nature Communications *11*, 735. 10.1038/s41467-019-13983-9.