**Supplementary information**

# 1. Supplementary notes

## 1.1 Pseudo-code for CrossICC procedure

**Procedure** CrossICC

**Input**: a set of data matrix from different batch/study/platform

a number of max iteration *H*.

a series of initial filter options *f(x)*, see **1.1.1** below

a cluster method applied in Consensus Clustering

**for**  **do**

{initialize data matrix }

**end** {for }

{store the origin feature names, }

**while** **and**  **do** {is the updated feature list after an iteration}

{matrix of centroids by clusters, initially empty}

{list that stores the sample cluster number label of matrix **D**, initially empty}

{set of feature list of matrix **D**, initially empty}

**for**  **do**

{subset **D’** with , skipped at the first run}

{clustering into clusters, the of each matrix were obtained by CDF < 0.05, is cluster labels of samples in }

{compute centroids of cluster in each by }

**end** {for }

{get supper clusters number of centroids matrix by hierarchical clustering, the optimize cluster number was obtained by silhouette analysis}

**for**  **do**

{assign the supper cluster number to each individual sample}

**end** {for *m*}

**for do**

get differentially expressed features list according to the in each ,see **1.1.2,** also known as MDEG

**end** {for *s*}

get consensus features **E** across all matrix **D**

**end** {while}

**return** ***K*** and ***e***

* + 1. **Filter steps of function** *f(x)*

In CrossICC, *f(x)* consist of three data preprocessing steps the make sure that the data matrix in each platform can be properly fed into iterations. First, only genes or features present in all platform were kept and MergeMaid were applied to remove low stability/ reliability ones. Second, genes with low variability that have MAD (median absolute deviation) < 0.5 were removed to accelerate the further calculation. Third, expression values within platform were normalized by median centered method. Note that users should decidedly discharge samples or even platforms if it has extremely unusual values or distribution, otherwise there might be few features or samples left in the initial filter step.

**1.1.2 Get differential expressed genes and MDEG**

In each iteration, samples from different platform were assigned to meta-clusters. Differential expression analysis was performed based on cluster labels within each platform. For a certain cluster, by using F-test (adjust P < 0.05 by default, this value could be customized to avoid no genes were retained), we retained the significantly upregulated gene in this cluster against others in the same platform. Similarly, the cluster specific genes were collected in each platform. We then filtered out the genes that have inconsistent expression tendency across platforms. To note, for a certain cluster, platforms that has very few samples in this cluster were not considered (less than 5 by default). Finally, given the genes were specifically expressed in each meta-cluster across platforms, we store the gene names into a to do comparison with previous run . The iteration stopped only the two set were exactly the same or it reach the defined max iterator time (20, default). When iteration stopped, set from the last round could be defined as MDEG.

**1.2 Single sample gene set enrichment analysis**

To biologically characterize each patient of the cancer subtypes, a single sample gene set enrichment analysis (ssGSEA) algorithm was implemented (6). ssGSEA will calculate the enrichment score of the biological pathways in each single sample. The details of algorithm are as follows:

The *N* genes of a given sample *S* were ranked and the rank was used to replace the expression value, and then a weighted value is assigned to each gene according to its rank. is calculated as follows:

equation 1

Then a Normalized Centroid is defined as the uni-dimensional average of the weighted rank values of the genes in a given gene-set , and is the uni-dimensional average of the weighted rank values of the remaining genes NG. and are calculated as follows:

equation 2

equation 3

(where |G| is the number of genes in gene set G , |NG| is the number of the genes not in gene set G).

The enrichment score is defined by calculating the difference between the normalized centroid of its gene-set and that of its complement gene-set. is calculated as follows:

equation 4

# 2. Supplementary Figures

PCAplot.pdf

**Figure S.** **Variance contribution of top 10 principal components in individual dataset and merged dataset**. “Total” denotes the merged dataset after batch effect correction.