a

CRC (C:396, N:45)
PDAC (C:261, N:25)
HCC (C:379, N:50)
EAC (C:125, N:64)
ESCC (C:97, N:80)
GC (C:395, N:23)

450K tissue data

↓

Tissue specific CpGs (30Mb)

↓

Targeted bisulfite sequencing of selected tissue CpGs in 300 plasma samples

Test set (30% samples) ←

↓

Model evaluation

Training set (70% samples)

↓

DMR calling and feature selection

↓

Random forest model training

Normal (46)
CRC (40)
PDAC (74)
HCC (43)
EAC (12)
ESCC (48)
GC (37)

b

Open sea (interCGI)
CGI   CGI shore
CGI shelf

Fig. S3. Workflow of model training and test for GI cancer prediction. (a) Initial markers discovery was conducted with 1653 GI cancer and 287 normal tissues. Targeted bisulfite sequencing that covers the initial selected CpG sites was performed on 300 plasma samples, including 254 GI cancer and 46 normal samples. GI plasma samples set was split into training set (70%) and test set (30%). The training set was used for calling differential methylated regions, feature selection and training random forest prediction models. The test set was used for model performance evaluation. (b) Annotations of panGI markers selected from GI cancer tissues.