we found the top 289 most importance regions could take account of 80.0% contribution to the accurate prediction.

Random forest algorithm showed 1585 regions could provide positive ability to distinguish cancer samples from normal samples while the top 286 most importance regions could take account of about 80.0% contribution, with sensitivity of 97.06%, specificity of 100% and accuracy of 97.37%.

The second round prediction process of random forest model based on top 206 regions with mtry of 14 and mtrees of 500 showed 100% sensitivity, 100 specificity and 100% accuracy. One the other side, 248 regions were hypermethylated in at least 50% cancer samples. The average length of the regions were 103bp (IQR=95bp, SD=126) and were located in the promoter region of 182 genes (2000bp up-stream of TSS). With the help of text mining, 21 genes of them were validated to be methylation relevant cancer related genes (Table).

Next-generation methylation sequencing and quality control

Unique mappable reads

In the first step,

In the next step, genome-wide DNA methylation profiles of 30 samples including 15 solid cancer tissues (5 colon cancer, 5 lung cancer and 5 pancreatic cancer) and corresponding plasmas detected with [reduced representation bisulfite sequencing](http://en.wikipedia.org/wiki/Reduced_representation_bisulfite_sequencing) (RRBS) were collected to discover the pattern of the shedding for the methylated DNA fragments from tissues to blood.

Task 1.

1, You need to filter all the hypermethylated fragement in cancer solid tissues and corresponding plasma circulating DNA while no methylation signals in health plasma.

2, and then validation these signals in TCGA database.

Task 2. Different method comparison.

Can it be used on RRBS data? We can downsample high coverage data to lower the effect of clonal reads on the analysis.

Noi to check the input requirement for performing RRBS, can we perform RRBS in parallel on the test samples.

The optimal number of variables tried at each split is 16 (mtry) and number of trees (mtrees) makes no difference to the prediction accuracy from 500 to 7000. RF prediction model showed 1585 regions could provide positive ability to distinguish cancer samples from normal samples with high specificity of 100%, however, the sensitivity was only 25%, which indicated large number low predictive biomarkers were enrolled into the prediction model. After removed the most 60% lower informative regions in the random forest model the sensitivity was only 78%, (Informative regions from BSPP see supplementary Table \*)

Therefore, we merged with the informative biomarkers identified in RRBS and Cap-seq dataset, eventual, there are 2 biomarkers were remained, including:

In the prediction section, top 62 regions could provide the distinguish accuracy of 86.87% which did not indicated Capseq method were worse than RRBS in the prediction ability. First capture, On the another side, when you will top 10 predictive regions, the separate ability of prediction model based on Capseq dataset could come up to the sensitivity of 90.32%, specificity of 95.83% and accuracy of 92.49%. However, small number of predictors would greatly decrease the robust or reproducibility of the prediction model and would bring biased inference in the process of clinical application of Capseq assay.

63.4% MHL distance of BSPP were less than 200bp, indicating they are almost located in the same region/CpG island in human genome, therefore, we need merge these MHL region together to increase the sensitivity of the prediction.