Supplementary Material

Methodology for Figure 2

The following approach was used to assess the applicability of cfDNA based methylation biomarkers on 450k methylation bulk samples. First, we manually curated a list of cfDNA methylation markers from the literature. Literature mining provided us 48 unique cfDNA methylation biomarkers for breast cancer. 450k methylation data-set for bulk tissue was retrieved from TCGA database (752 breast cancer, 96 non-cancerous) and processed for markers based overlapping regions using R package "GenomicRanges". To demonstrate the utility of single marker for classification, a balanced dataset of 192 bulk (TCGA) 450k methylation samples was created based on random selection (96 breast cancer, 96 normal). Later the TCGA data was used for randomly selected biomarker for predicting sample label (cancer vs non-cancer) . Based on LDA fitting, values for sensitivity and FPR were obtained and presented in the form of box-plots [Figure 1(a)]. Secondly, the diagnostic potential of breast cancer cfDNA methylation marker set (literature derived) on 848 TCGA samples was inspected. Markers based normalised beta scores for all the TCGA observations were visualised as heatmap for differential analysis [Figure 1(b)].

Methodology for Figure 3

To study the potential of cfDNA 5hmC methylation as a biomarker in predicting the cancer types, we used machine learning methods. Using the 5hmC profiles published by Song et al., we first estimated read-count on RefSeq genes and CPG islands downloaded from the UCSC genome browser. Using the 5hmC read-count of cfDNA on RefSeq genes we applied dimension reduction using tSNE to visualise separability among samples from different types of disorder. We repeated the same procedure using all CPG island. Later we choose top 50 Refseq genes and top 50 CPG islands. For this purpose, we choose top 50 genes with good feature importance provided by random-forest method. However, before applying random-forest we performed upsampling of data so that the number of samples from all disorders remains the same. For upsampling we added Gaussian noise to cfDNA 5hmC read-counts so that every upsampled profile is different from each other. After upsampling and equalising the number of samples from each disorder, random forest was applied and top 50 predictors were chosen. The signal (read-count) on top 50 predictors (CPG islands or genes) were then used for dimension reduction using tSNE. Notice that upsampling was performed only to choose top 50 features (markers). Dimension reduction was done for original number of samples. After dimension reduction using tSNE, we applied DBSCAN to cluster the samples. We calculated clustering purity using normalised mutual information. We used R programming language for the application of tSNE and DBSCAN and the application of NMI. Default value of tSNE parameters were used here.

Methodology for Figure 4

In order to understand the use of deconvolution algorithms for distinguishing cancerous samples from non-cancerous, we exhibited the performance of three most commonly used deconvolution techniques i.e. RefFreeEWAS, ReFACTor and SVA on bulk (TCGA) and cfDNA methylation samples.

Firstly, methylation beta values corresponding to 100 randomly selected CpG islands were retrieved for 100 TCGA 450k methylation samples (50 prostate cancer,50 normal) from TCGA database. Since the three mentioned approaches for deconvolution belong to reference-free category, the source estimation was done based on scree plot. The number of factors to be retained was chosen to be 5, as per the scree analysis. Subsequently tSNE was applied on the normalised TCGA data as well as on respective deconvoluted matrices (perplexity=5, max_iter = 300) [Figure 4(a)]. Similar workflow was followed for cfDNA 450k methylation samples as obtained from CFEA database (14 prostate cancer,14 normal). The estimated value for cell types was 3 and tSNE parameters were set as perplexity=3, max_iter = 60 [Figure 4(b)].