**Supplementary Methods**

1. **Leave-one-out analysis**

The leave-one-out procedure was performed according to the following steps. For a sample size of n, we left one sample out as a testing sample, and used the remaining samples (n – 1) to train the classifier based on support vector machine (SVM) and logistic regression using the 256 plasma DNA end motifs. Then, we used the trained classifier to determine whether the left-out sample was classified as taken from a subject with or without cancer. We systematically left one sample out as a testing sample to test the classifier trained from the remaining samples. Therefore, we could obtain a predicted result for each sample and the accuracy was calculated from the predicted results.

1. **Plasma DNA end motif definition.**

Plasma DNA libraries were analyzed by paired-end massively parallel sequencing. The sequenced fragments were aligned to a human reference genome. The first 4-nucleotide (nt) sequence (i.e. a 4-mer motif) on each 5’ fragment end of plasma DNA was determined in relation to the reference genome. Using a 4-mer motif, there would be a total of 256 possible motifs (i.e., 44). The frequency of occurrence of each motif was calculated and normalized by the total number of ends among the sequenced fragments. As illustrated in Figure 1, during library preparation, the end-repair step would modify the original 3’ ends by digesting the 3’ protruding single-stranded ends or filling in the 3’ receded ends with the use of 5’->3’ polymerization activity and 3’->5’ exonuclease activity of the Klenow fragment of DNA polymerase. Therefore, it was expected that all molecules would be repaired to become blunt-ended and reflect the original 5’ ends of each complementary pair of DNA fragments, while the 3’ end motifs would be modified. Thus, for the downstream analysis using aligned fragments, the 5’ end motifs on both sides of each double-stranded plasma DNA molecule were analyzed together to reflect the true profile of plasma DNA end motifs (Fig. 1 of the main text).

Table S1 and S2 contained the raw data regarding the end motif frequencies.

1. **Computer simulation**

The plasma DNA of a patient with cancer comprised the tumor-derived DNA (with a tumor DNA fraction ***f***) and the non-tumor-derived background DNA (1-***f***). The plasma DNA end motifs frequency (***Mp***) was assumed to be a linear combination of the corresponding tumoral and nontumoral DNA end motif distributions, namely, ***Mt*** and ***Mn***, which were weighted by the tumoral and non-tumoral DNA fractions by the following formula:

$f×M\_{t}+(1-f)×M\_{n}=M\_{p}$**, (1)**

where $M\_{n}$was generated by the motif distributions of plasma DNA fragments carrying wildtype alleles based on Dirichlet distribution; $M\_{t}$was generated by the motif distributions of plasma DNA fragments carrying mutant alleles based on Dirichlet distribution. $M\_{n}$and$M\_{t}$would be governed by formulas (2) and (3) below**:**

$M\_{n}\~ rdirichlet(WT\*W)$ **, (2)**

$M\_{t}\~ rdirichlet(MT\*M)$ **, (3)**

where ‘***rdirichlet****’* was the function to generate data from the Dirichlet distribution which was implemented in a R package, MCMCpack (https://cran.r-project.org/web/packages/MCMCpack/index.html); ‘***WT***’ represented the end motif frequencies of those fragments carrying wildtype alleles; ‘***MT***’ represented the end motif frequencies of those fragments carrying wildtype alleles; ‘***W***’ represented the number of fragments carrying wildtype alleles; ‘***M***’ represented the number of fragments carrying mutant alleles.

The number of plasma DNA molecules was denoted by ***R***. The counts of end motifs drawn from nontumoral end motif distribution ($Nontumoral\_{end motifs}$) followed the multinomial distributions below:

$Nontumoral\_{end motifs}\~ rmultinom\left(R×\left(1-f\right), M\_{n} \right)$**, (4)**

The counts of end motifs drawn from tumoral end motif distribution ($Tumoral\_{end motifs}$) followed the multinomial distributions below:

$Tumoral\_{end motifs}\~ rmultinom\left(R×f, M\_{t} \right)$**, (5)**

where ‘***rmultinom***’ was the multinomial function which was implemented in R package, MCMCpack.

$Nontumoral\_{end motifs}$and$Tumoral\_{end motifs}$were combined to calculate the motif diversity score (MDS).

We studied the discriminating power of detecting patients with HCC using MDS. To this end, we simulated the tumor DNA factions (***f***) with 0.1%, 1%, 2%, 3%, 4%, 5%, 10%, 20%, and 30%, respectively. At each tumor DNA fraction (***f***), we simulated the effect of varying the number of plasma DNA molecules encompassing 500, 1000, 5000, 10,000 50,000, 100,000, 500,000, 1,000,000, 2,000,000, 5,000,000, 10,000,000, 20,000,000, and 30,000,000, assuming tumor DNA fractions (***f***) of 10% and 5%, separately. For each tumor DNA fraction ***f*** with a given amount of plasma DNA molecules, we simulated MDS values for 100 subjects with HCC and 100 subjects without HCC, respectively. We determined the area under the ROC curve between patients with and without HCC using MDS values.

**Statistical analysis and heatmap.** The ROC curves and heatmap were constructed using the open source statistical R package pROC and gplots (R package), respectively. P-values involved 256 plasma DNA end motifs were adjusted by the Bonferroni procedure because of multiple comparisons.