Supplementary Information

Supplementary materials and methods

Ethics statement

We established the Community-based Cohort study on Population with high risk of Liver Cancer (the CCOP-LC cohort; Chinese Clinical Registry, ChiCTR-EOC-17012853) in 2017, based on the early HCC screening program conducted in the community population. The study protocol (NCC201709011) was approved by the Institutional Review Board of National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences in Beijing, China.

Overview of early HCC screening program in community populations

Early HCC screening was conducted according to the "Technology Scheme of Early Diagnosis and Early Treatment for Cancer" issued by the China Expert Committee of Early Detection and Early Treatment of Cancer, The Ministry of Health Bureau of Disease Prevention and Control (1). A population-based cancer registry and department of vital statistics have been established in all of the screening centers (2). Briefly, HBsAg-positive "healthy" individuals 35-69 years old were invited to participate in early HCC screening. All participants underwent testing for determination of serum AFP concentrations and ultrasonography (US; Aloka ProSound SSD-4000; Shanghai, China), as well as other standard biochemical tests (Table 1). Individuals were designated as AFP/US- positive, suspected, or negative based on AFP serum levels and the detection of liver nodules. "AFP/US-positive" individuals had any of the following: 1) serum APF levels of > 400 ng/mL regardless of US-detected nodule; 2) US-detected nodule of ≧ 2 cm in size regardless of serum AFP concentration; 3) US-detected nodule of ≧ 1 cm in size with serum AFP ≧ 200 ng/mL. The "AFP/US-suspected" individuals had either of the following: 1) serum AFP levels of ≧ 20 ng/mL regardless of US-detected liver nodule; 2) US-detected nodule of ≧ 1 cm in size. The "AFP/US-negative" individuals were defined as having serum AFP levels of < 20 ng/mL without a US-detected liver nodule. The AFP/US-positive individuals were referred to advanced hospitals (a level 3 hospital in China) for confirmation diagnosis as determined with dynamic CT or dynamic MRI, and they received relevant therapy based on clinical practice guidelines (Fig. 1)(3). Individuals without confirmation diagnosis were invited to return to undergo dynamic CT/MRI in 2 months. AFP/US-suspected individuals were recommended to receive second round examination for serum AFP quantification and US in 2-3 months.

Participants and study design

The participants in the current study were derived from the CCOP-LC cohort based on individuals evaluated at four screening centers from Jiangsu and Anhui provinces in China (Fig. 1). During the AFP/US screening (considered baseline, conducted between Oct 7, 2017 and January 31, 2018), we collected peripheral blood (5 mL in EDTA coated tubes), which was centrifuged within 2 h of collection at 4000×*g* for 10 min to separate plasma and blood cells. All samples were stored at -80°C. In most cases, 0.5 mL plasma was used to determine protein markers and 2 mL plasma for cfDNA extraction.

The 176 AFP/US-positive/suspected cases were further analyzed in the HCCscreen assay. According to the diagnosis in the follow up examinations, the participants with reliable diagnoses were selected as the training cohort in this study. To validate our findings, we sampled 331 participants from the AFP/US-negative individuals with similar age to those who were AFP/US-positive/suspected in the HCCscreen assay. From May 20 to July 17, 2018 (6-8 months after the baseline blood draw), the 331 individuals were followed by offering an examination of dynamic CT/MRI, AFP/US or telephone interview. The CT/MRI images were independently evaluated by two radiologists from the National Cancer Center, Chinese Academy of Medical Sciences in Beijing. During this period, we offered an additional AFP/US test to the individuals who were AFP/US-negative at baseline and had not taken the HCCscreen test. Some of them did not choose the additional AFP/US examination, and their liver cancer outcome (ICD-10 code C22) by June 30, 2018 was obtained from the population-based cancer registry in the screening centers (Fig. 1). Of the 3617 AFP/US-negative individuals, 1612 (44.6%) participants were able to be followed during May 20 to July 17, 2018, which was 6-8 months after baseline screening. Among them, 87 participants received dynamic CT/MRI examinations, 1120 received AFP/US, and 68 were interviewed by telephone. The liver cancer outcome in 337 participants was obtained from local population-based cancer registry (*SI Appendix*, Fig. S1). The HCC status in the other 2005 participants was not available by June 30, 2018 (*SI Appendix*, Fig. S1).

The 70 healthy controls were derived from those who did the annual physical examination and did not report any HBV infection. All were confirmed as HBsAg-negative when blood was donated.

Determination of serum DCP concentrations

Serum DCP levels were determined using a commercialized kit in the Abbott ARCHITECT i2000SR Chemical luminescence immunity analyzer (CLIA) according to the manufacturer's instructions (Abbott Laboratories; Chicago, IL, USA).

Profiling of cfDNA alterations

We designed an assay to sequence cfDNA to profile: 1) the coding regions of *TP53*, *CTNNB1*, *AXIN1* and the promoter region of *TERT* (Table S1); 2) HBV integrations. Briefly, cfDNA fragments were first ligated to an adaptor with random DNA barcodes (*SI Appendix*, Fig. S2). The ligated constructs were amplified through 10 reaction cycles to generate a whole genome library, containing hundreds of redundant constructs with unique DNA barcodes identifying each original cfDNA fragment. The amplified library was sufficient for 5-10 independent sequencing analyses. The target regions were amplified together with the DNA barcode in 9 cycles of PCR using a target-specific primer (TS primer 1) and a primer matching the adapter sequence (4, 5) (*SI Appendix*, Fig. S2). A second round of 15 cycles of PCR with one pair of nested primers matching the adapter and the target region (TS primer 2) was used to further enrich the target region and add the Illumina sequencing adapter (*SI Appendix*, Fig. S2). An efficient enrichment was observed in the PCR based assay with > 80% of reads mapping to a small target region of < 10 Kb. Using this assay, we can cover the target region > 100,000 times with 3 Gb of sequencing data, enabling 20× redundant sequencing for 5,000 copies of original cfDNA. With the DNA barcode ligated to the original cfDNA molecule, we can track redundant reads from an original cfDNA molecule to minimize calling errors inherent in PCR amplification and parallel sequencing as mutations(6, 7). We examined 11 mutations detected by this assay with digital PCR and validated all these mutations with a range of 0.03-0.16% mutation fractions.

Data process and mutation detection

Sequencing reads were primarily processed with our own program to extract tags and remove sequence adapters. Residual adapters and low-quality regions were subsequently removed using Trimmomatic (v0.36). The cleaned reads were mapped to the hg19 and HBV genomes using 'bwa (v0.7.10) mem' (8) with the default parameters. Candidate somatic mutations, consisting of SNP and INDEL, were identified using samtools mpileup (9) across the targeted regions of interest. To ensure accuracy, reads with the same tags, and start and end coordinates were grouped into Unique Identifier families (UID families). UID families containing at least two reads and in which at least 80% of reads were the same type were defined as Effective Unique Identifier families (EUID families). Each mutation frequency was calculated by dividing the number of alternative EUID families by the sum of alternative and reference ones. The mutations were further manually reviewed in IGV. The candidate variations were annotated with Ensembl Variant Effect Predictor (VEP) (10). HBV integrations were identified using Crest (11) , and at least 4 soft-clip reads supports were needed.

Predictive model construction

1. Feature mapping and data preprocessing

1) Mutation annotations and scoring:

Considering that several factors may be functionally related to the penetration of mutations, and in turn be related to the phenotypes, we weighted each mutation with two relevant factors and scored each with a calibrated linear combination of factors.

- <u>PAPI score annotation</u>

  PAPI is a machine learning ensemble algorithm to evaluate the functional penetration of human DNA mutations (12). It is comprised of two popular annotation algorithms, VEP and POLYPGEN, and provides an accurate score of deleteriousness of mutations.

- <u>Annotation of frequency of mutations</u>

  Mutation frequency (the fraction of reads support for a candidate mutation) is highly proportional to the overall quantity of circular tumor DNAs in the blood as well as the tumor size. Therefore, we annotated all the input mutations with their reads support frequencies.

- <u>Structural variants as features</u>

  Another two features of structure variance have been used to construct the model, including TERT fusion and counts for the HBV fusion events.

2) Collapsing of mutations

Several genetic features were extracted by collapsing the mutations into either gene-level or focal regions. For each region of interest (ROI), the ROI score was calculated by

$$\mathrm{ROI} = \log_2 \sum_{i=1}^{n} adj_{score}$$

where n is the number of mutations overlapping the ROI, and the adj_score is the weighted sum of above mentioned mutational factors, the PAPI and frequency of reads support score. The weighted vector was adjusted so that the model performance was maximized.

3) Protein and experimental markers

Two protein markers, DCP and AFP, were used in our model, as they have been shown in previous studies to be very strong indicators for HCC diagnosis (13). The values were ranked into several classes of numeric values. The ctDNA concentration was included in our model feature list as well.

4) Clinical information as features

The age and gender of a given patient were built into our predictor as well, as it has been demonstrated that the probability of HCC diagnosis is to some extent related to the age and gender of the individual.

2. Feature selection

RandomForest was applied to screen useful variables from the candidates; we applied a backward variables subtraction by minimizing the unbiased out-of-bag error estimation, with one feature eliminated at each run. Markers including proteins, genetic variants as well as clinical information, were then optimized as the final features to build a binary classifier. In the training of HCC vs healthy individuals, only ctDNA SNP/indel mutations and protein markers were used. HBV-TERT fusion or other HBV integrations were not included as the healthy group was free of HBV infection.

3. Model and parameters optimization

A logistic regression model was constructed from the training cohort of 135 samples which includes 65 HCC and 70 non-HCC. The model performance was evaluated both on the training and validation data sets, by the Area Under Curve (AUC) statistics. Sensitivity and specificity of the model was also determined using an optimized cut-off value of 0.4. This cut-off value optimization was applied using Youden's index. Cross-validated coefficients for each feature using logistic regression have been given as well, for the purpose of clustering analysis for genetic, protein and CNV levels, respectively. The model was initiated in R package 'glmnet' (R version 3.5.1), and the penalty parameter alpha was optimized with 10-fold cross validation within the training data set and the optimized value was 0.

Statistical analysis

We used a logistic regression model with ctDNA mutations, protein biomarker levels as well as clinical features as variables. We defined HCC cases and non-HCC cases with dynamic CT/MRI and/or histology in the AFP/US-positive and AFP/US-suspected individuals (Fig. 1). With 100 iterations of LOOCV (Leave-One-Out Cross Validation) on a training data set of 65 HCC and 70 non-HCC cases, we calculated the sensitivity and specificity of the HCCscreen assay.

References

1. Shia YC, Beever JE, Lewin HA, & Schook LB (1991) Restriction fragment length polymorphisms at the

porcine t complex polypeptide 1 (TCP1) locus. *Anim Genet* 22(2):194.

2. Chen W, *et al.* (2018) Cancer incidence and mortality in China, 2014. *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu* 30(1):1-12.
3. Omata M, *et al.* (2017) Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. *Hepatol Int* 11(4):317-370.
4. Perera BP & Kim J (2016) Next-generation sequencing-based 5' rapid amplification of cDNA ends for alternative promoters. *Analytical biochemistry* 494:82-84.
5. Zheng Z, *et al.* (2014) Anchored multiplex PCR for targeted next-generation sequencing. *Nature medicine* 20(12):1479-1484.
6. Kinde I, Wu J, Papadopoulos N, Kinzler KW, & Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 108(23):9530-9535.
7. Chaudhuri AA, *et al.* (2017) Early Detection of Molecular Residual Disease in Localized Lung Cancer by Circulating Tumor DNA Profiling. *Cancer discovery* 7(12):1394-1403.
8. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
9. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
10. Wang J, *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 8(8):652-654.
11. McLaren W, *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome biology* 17(1):122.
12. Limongelli I, Marini S, & Bellazzi R (2015) PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC bioinformatics* 16:123.
13. Chen H, *et al.* (2018) Direct comparison of five serum biomarkers in early diagnosis of hepatocellular carcinoma. *Cancer management and research* 10:1947-1958.

Fig. S1

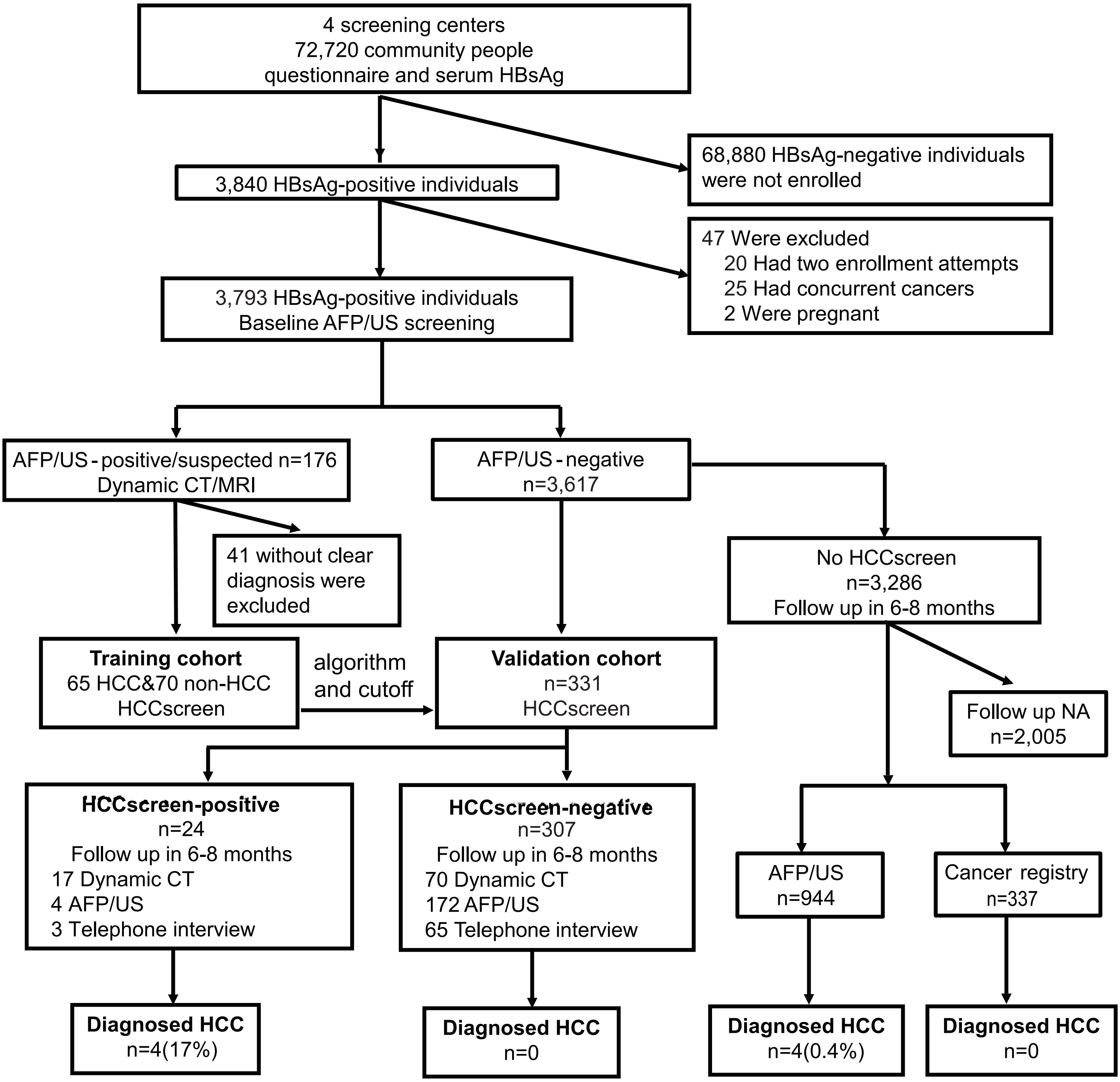

Fig. S1. Detailed study design.

Fig. S2

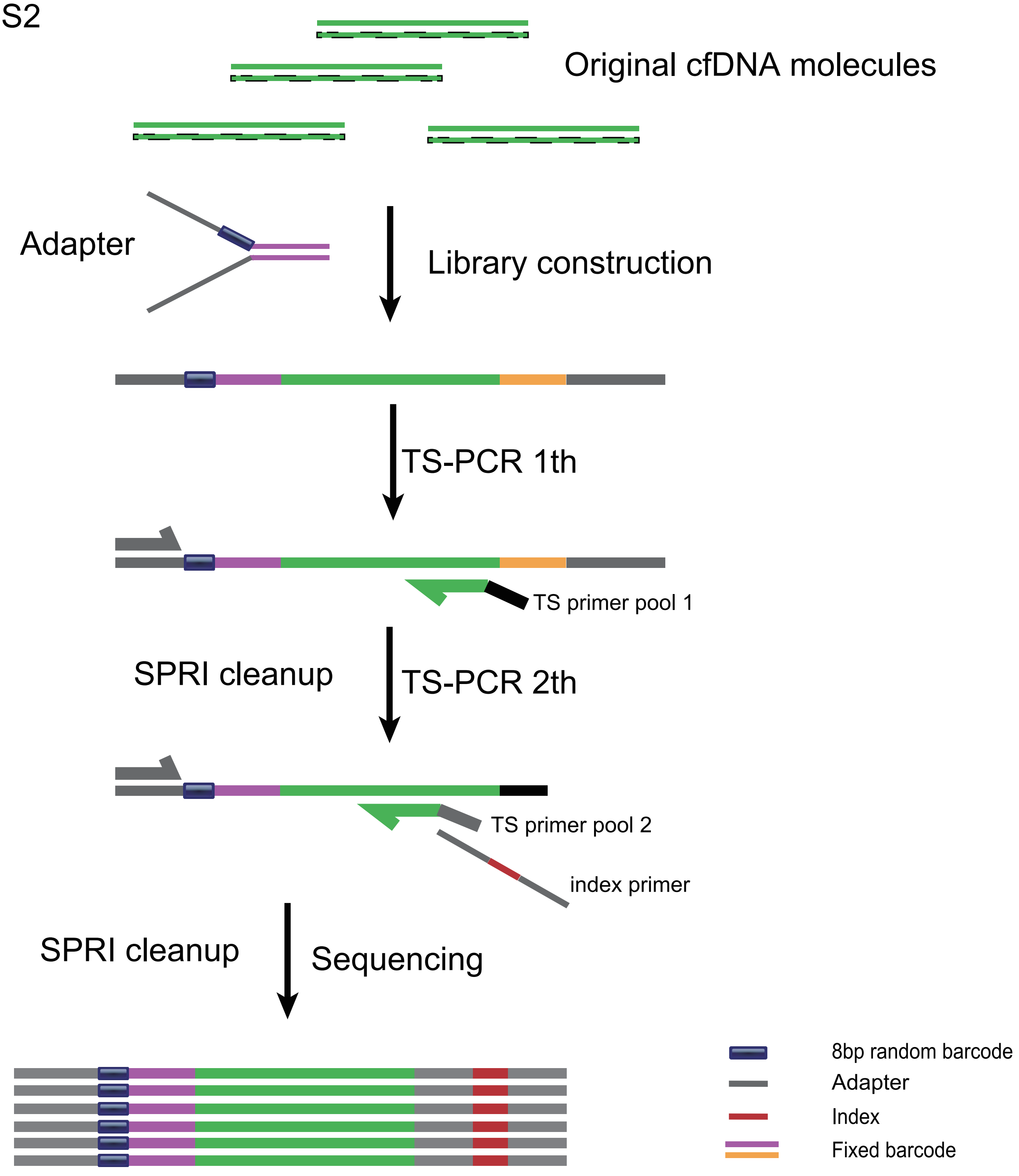

Fig. S2. The design of the genetic profiling of cfDNA in the HCCscreen assay.

Fig. S3

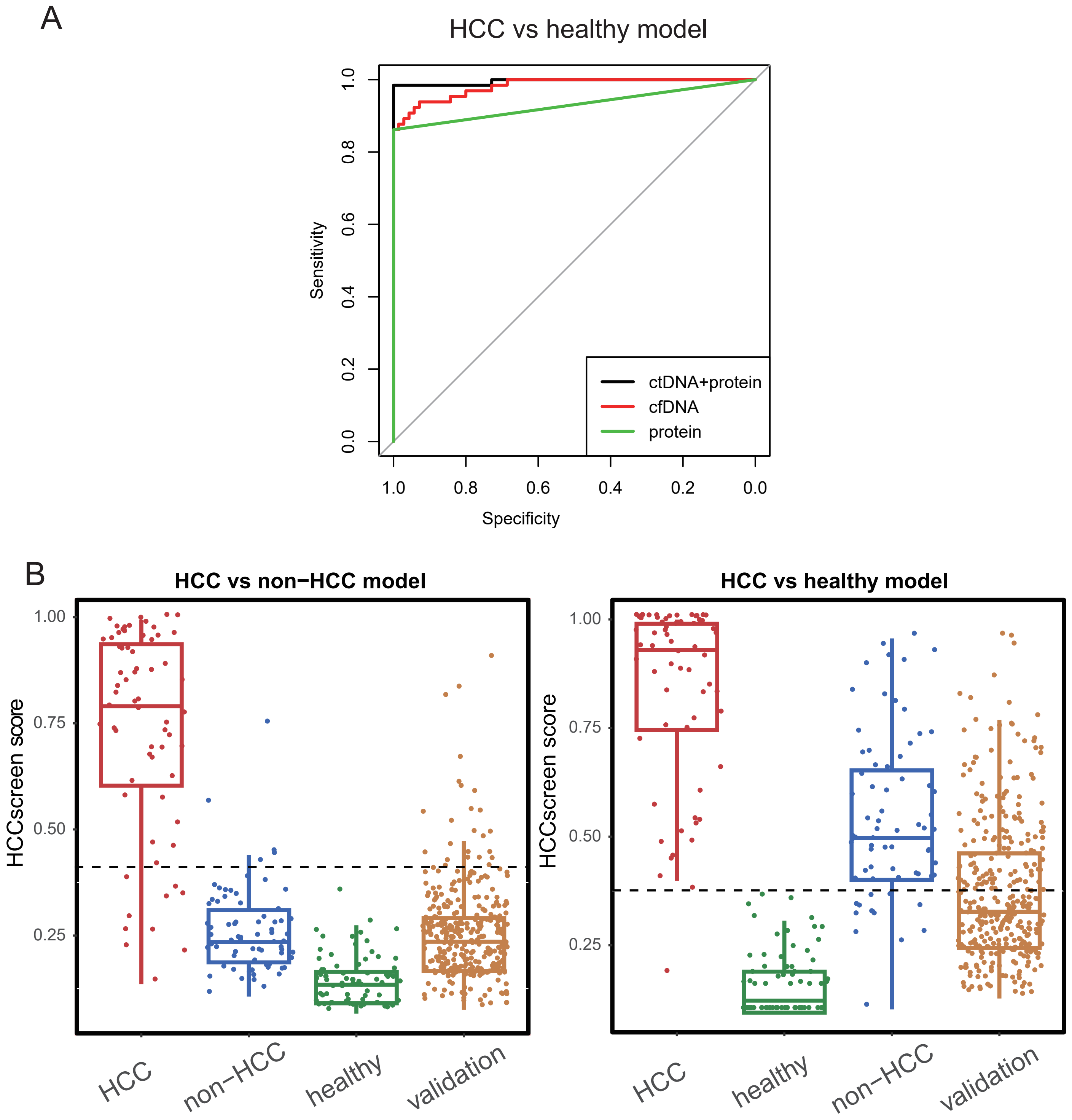

Fig. S3. The performance of different training cohorts.
A. ROC of the diagnostic model of the HCCscreen assay in the training cohort using healthy individuals without HBV infection as controls. **B.** Training with HCC and non-HCC individuals (left) and training with HCC and healthy individuals (right).