Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma

Peiyong Jiang[a,b,1], Kun Sun[a,b,1], Yu K. Tong[a,b], Suk Hang Cheng[a,b], Timothy H. T. Cheng[a,b], Macy M. S. Heung[a,b], John Wong[c], Vincent W. S. Wong[d,e], Henry L. Y. Chan[d,e], K. C. Allen Chan[a,b,f], Y. M. Dennis Lo[a,b,f,2], and Rossa W. K. Chiu[a,b,2]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [b]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [c]Department of Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [d]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; [e]Institute of Digestive Diseases, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; and [f]State Key Laboratory in Translational Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China

Contributed by Y. M. Dennis Lo, September 24, 2018 (sent for review August 28, 2018; reviewed by Ellen Heitzer and David Sidransky)

Circulating tumor-derived cell-free DNA (ctDNA) analysis offers an attractive noninvasive means for detection and monitoring of cancers. Evidence for the presence of cancer is dependent on the ability to detect features in the peripheral circulation that are deemed as cancer-associated. We explored approaches to improve the chance of detecting the presence of cancer based on sequence information present on ctDNA molecules. We developed an approach to detect the total pool of somatic mutations. We then investigated if there existed a class of ctDNA signature in the form of preferred plasma DNA end coordinates. Cell-free DNA fragmentation is a nonrandom process. Using plasma samples obtained from liver transplant recipients, we showed that liver contributed cell-free DNA molecules ended more frequently at certain genomic coordinates than the nonliver-derived molecules. The abundance of plasma DNA molecules with these liver-associated ends correlated with the liver DNA fractions in the plasma samples. Studying the DNA end characteristics in plasma of patients with hepatocellular carcinoma and chronic hepatitis B, we showed that there were millions of tumor-associated plasma DNA end coordinates in the genome. Abundance of plasma DNA molecules with tumor-associated DNA ends correlated with the tumor DNA fractions even in plasma samples of hepatocellular carcinoma patients that were subjected to shallow-depth sequencing analysis. Plasma DNA end coordinates may therefore serve as hallmarks of ctDNA that could be sampled readily and, hence, may improve the cost-effectiveness of liquid biopsy assessment.

tumor-associated preferred ends | liver-associated preferred ends | tumor-derived cell-free DNA | hepatocellular carcinoma | transplantation

DNA molecules released by malignant cells are present in the peripheral circulation of cancer patients, providing noninvasive access to such genetic material. Circulating tumor-derived cell-free DNA (ctDNA) analysis has been utilized as a liquid biopsy for the management of cancer. Liquid biopsies may serve as a surrogate of invasive biopsies because ctDNA molecules harbor molecular features that are associated with cancers. ctDNA features that have been characterized include somatic mutations, cancer-associated viral sequences, copy number aberrations, and differential DNA methylation signatures (1–5). We envision that the most direct means to detect cancer using a cell-free DNA sample may be through the detection of cancer-associated hallmarks that are physically present on cell-free DNA molecules. Such DNA molecules would be deemed as informative signals for the presence of cancer. Among the methods mentioned above, one disadvantage of approaches based on the detection of viral nucleic acids is that not all cancers are associated with viral infections. However, the use of DNA methylation analyses require performance of additional laboratory steps such as bisulfite conversion. As a result, many research groups have focused on the detection of somatic mutations in plasma (4, 6–8).

For the purpose of early cancer detection (1, 4, 9, 10), testing approaches need to be able to detect biomarkers that are broadly represented among the majority of cancer cases in the target population without a priori information of the tumor genetic

Significance

Cell-free DNA fragmentation is a nonrandom process. We showed that cell-free DNA fragments with ends at certain genomic coordinates had higher likelihoods of being derived from hepatocellular carcinoma. Other coordinates were associated with cell-free DNA molecules originating from the liver. Quantitative assessment of cell-free DNA molecules bearing these respective groups of end signatures correlated with the amounts of tumor-derived or liver-derived DNA in plasma. There were millions of tumor-associated plasma DNA end coordinates across the genome. Due to their high prevalence, they were more readily detectable than somatic mutations as a cancer signature in plasma. Hence, detection of tumor-associated plasma DNA ends may offer a cost-effective means of capturing evidence for the presence of cancer through liquid biopsy assessment.

Author contributions: K.C.A.C., Y.M.D.L., and R.W.K.C. designed research; P.J., K.S., Y.K.T., S.H.C., T.H.T.C., M.M.S.H., K.C.A.C., Y.M.D.L., and R.W.K.C. performed research; P.J., K.S., J.W., V.W.S.W., H.L.Y.C., K.C.A.C., Y.M.D.L., and R.W.K.C. analyzed data; P.J., Y.M.D.L., and R.W.K.C. wrote the paper; and T.H.T.C., J.W., V.W.S.W., and H.L.Y.C. recruited subjects and analyzed clinical data.

Reviewers: E.H., Medical University of Graz; and D.S., Johns Hopkins University School of Medicine.

Conflict of interest statement: The authors declare a conflict of interest. K.C.A.C., R.W.K.C., and Y.M.D.L. hold equities in Grail. P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. are consultants to Grail. K.C.A.C., R.W.K.C., and Y.M.D.L. receive research funding from Grail. Y.M.D.L. is a scientific cofounder of and serves on the scientific advisory board of Grail. P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. have filed patent applications based on the data generated from this work. Patent royalties are received from Grail, Illumina, Sequenom, and Xcelom.

Data deposition: Sequence data for the subjects studied in this work who had consented to data archiving have been deposited at the European Genome-Phenome Archive (EGA), https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (EBI) (accession no. EGAS00001003160).

[1]P.J. and K.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: loym@cuhk.edu.hk or rossachiu@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814616115/-/DCSupplemental.

Published online October 29, 2018.

profile. However, cancers are highly heterogeneous (11, 12). Even when a wide spectrum of mutations is targeted, not every cancer would bear at least one of those target mutations (4, 12, 13). Lawrence et al. (12) surveyed the total burden of somatic mutations in a broad range of malignancies. The number of somatic mutations per cancer genome ranged from ∼3,000 to 30,000. The spectrum of mutations observed include those known as driver mutations, but the majority of such mutations are nondriver in nature, with many of which being private to an individual tumor. ctDNA molecules circulate among a background of nontumor-derived cell-free DNA molecules. At earlier stages of cancer, the tumor DNA fractions in the circulation are typically lower than those amounts present at later stages of cancer. Further compounded by the issues of intratumoral heterogeneity, the fractional concentration of any one somatic mutation originating from only one of the subclones might be even lower in plasma (2, 7, 8).

To overcome these issues, we explored methods to improve the detection of informative ctDNA molecules in plasma. First, we developed an approach to detect the total pool of cancer-associated somatic mutations in plasma. Second, we investigated if there might be other classes of molecular signatures present on ctDNA. Recently, it has been shown that plasma DNA fragmentation is a nonrandom process (14, 15). In the plasma of pregnant women, there exists plasma DNA molecules with ends that are preferentially derived from the fetal tissues while other molecules contained ends that are preferentially derived from the maternal tissues (14). We hypothesized that tissue or tumor-associated preferred plasma DNA end signatures existed. We studied and compared the abundance of somatic variants and tumor-associated preferred ends as identifiable features of ctDNA.

Results

Clinical Specimens. This study was approved by the Joint Chinese University of Hong Kong and New Territories East Cluster Clinical Research Ethics Committee. Research participants were recruited from the Prince of Wales Hospital, Hong Kong, with informed consent. Paired plasma and tissue analysis was performed on a 60-year-old Chinese male who had chronic hepatitis B virus (HBV) infection complicated with compensated cirrhosis and hepatocellular carcinoma (HCC). Peripheral blood was collected before surgery. Tumor tissue and a biopsy of the adjacent nontumorous liver were collected at the time of resection. The tumor measured 18 cm at its largest dimension and was moderately differentiated without distant metastasis. Peripheral blood samples collected into EDTA tubes were also obtained from a chronic HBV carrier without liver cirrhosis nor HCC, 14 liver transplant recipients, and the corresponding liver donors. The liver transplant recipients were in stable conditions for more than 10 years after transplantation. Cord blood was collected from an uneventful Cesarean section delivery at 38 wk of gestation. Previously published sequencing data (16) from 32 healthy subjects, 67 chronic HBV carriers without cirrhosis, 36 patients with HBV-related liver cirrhosis, and 90 patients with HCC were also analyzed to investigate the tumor-associated cell-free DNA preferred ends.

Genomewide Scanning for Circulating Tumor-Derived Cell-Free DNA Signatures. Broad and deep survey of cell-free DNA was conducted by performing massively parallel sequencing on plasma DNA libraries prepared without PCR enrichment (14) until all of the library contents were consumed. The cell-free DNA sequence data after exhausting the sequencing library of one HCC patient reached >200× haploid human genome coverage. We then identified and quantified somatic mutations and tissue- or tumor-associated preferred DNA end signatures within the samples.

Somatic Mutation Identification from Tumor Biopsy. The buffy coat, resected tumor, adjacent normal liver tissue, and plasma specimen obtained from the HCC patient were sequenced to 45×, 45×, 40× and 220× human haploid genome coverage, respectively. The sequenced reads were first aligned using the SOAP2 aligner (17). Amounts of reads aligning to chromosomal segments were determined as described (2) and revealed copy number gains and losses in the tumor tissue biopsy and plasma but not in the normal liver tissue (*SI Appendix*, Fig. S1). We then aimed to identify sequence variants that were present in the HCC tumor tissue but not among the buffy coat sequencing data of the HCC patient. Initial analysis revealed 11,086,153 such variants. Using a strategy we have developed—termed "dynamic threshold filtering"—to remove apparent variants arising due to sequencing errors, 16,027 variants were retained. Sequence reads spanning these 16,027 putative variants were realigned using a second aligner, namely Bowtie2 (18). Reads carrying 12,112 putative variants were mapped to the same genomic location by both aligners. Lastly, 11,903 of the 12,112 putative variants were present at >30% fractional concentration.

To assess the false-positive rate of this filtering strategy, the same process was applied to the adjacent normal liver tissue biopsy. After dynamic threshold filtering, there were only 3,688 putative variants not present in the corresponding buffy coat DNA. Four hundred fifty-one of these variants remained after the second alignment step and only 96 of the 451 variants were present at >30% fractional concentration. To assess if the final sets of variants were true somatic mutations, we assessed the size of the DNA fragments bearing those variants in the corresponding plasma sample. *SI Appendix*, Fig. S2*A* shows that the cell-free DNA molecules bearing the variants identified from the tumor biopsy were shorter than the background cell-free DNA molecules in the plasma of this HCC patient. On the contrary, the 96 variants from the adjacent normal liver tissue did not show size shortening (*SI Appendix*, Fig. S2*B*). Tumor-derived cell-free DNA molecules have been shown to be shorter than the nontumor-derived molecules (16). Thus, the 96 variants from the normal liver biopsy were likely to be false positives. The 11,903 variants identified from the HCC biopsy were deemed true somatic mutations, and this number was compatible with the reported average number of mutations present per tumor (12).

Scanning for Somatic Mutations in Plasma Without a Priori Tumor Information. A depth of 220× haploid coverage was achieved after sequencing the entire PCR-free library prepared from the 4-mL plasma sample of the HCC patient. Bioinformatics filtering was performed without a priori knowledge of the set of somatic mutations identified from the tumor biopsy. Dynamic threshold filtering was applied to screen for sequence variants in both the plasma and buffy coat specimens. A total of 113,654 putative variants were present in the plasma but not the buffy coat. This set contained 10,050 (84%) of the 11,903 somatic mutations found in the tumor biopsy (Fig. 1*A*). After realignment, 11,504 variants remained. Putative variants present at >5% concentration amounted to 11,466 and included 80% of the somatic mutations. Lastly, 9,928 of the identified variants were present on sequence reads that were at least 10 bp shorter than the median read length of molecules carrying wild-type alleles in the sample and included 71% of the tumor-derived somatic mutations (16).

At each step of the filtering, variants that were present in the set of somatic mutations identified from the tumor biopsy were deemed true positives while those variants that were not within the set were classified as false positives. Based on these values, the positive predictive value (PPV) at each filtering step was determined. The PPVs improved from 8.8 to 85% after successive application of the different filtering steps (Fig. 1*A*). The 11,466 plasma DNA variants identified after the successive filtering steps except the size-based filter also showed a short size profile (Fig. 1*B*), consistent with the expected profile of ctDNA.

A cord blood plasma sample was sequenced to 100× using an identical PCR-free library preparation protocol. To test the

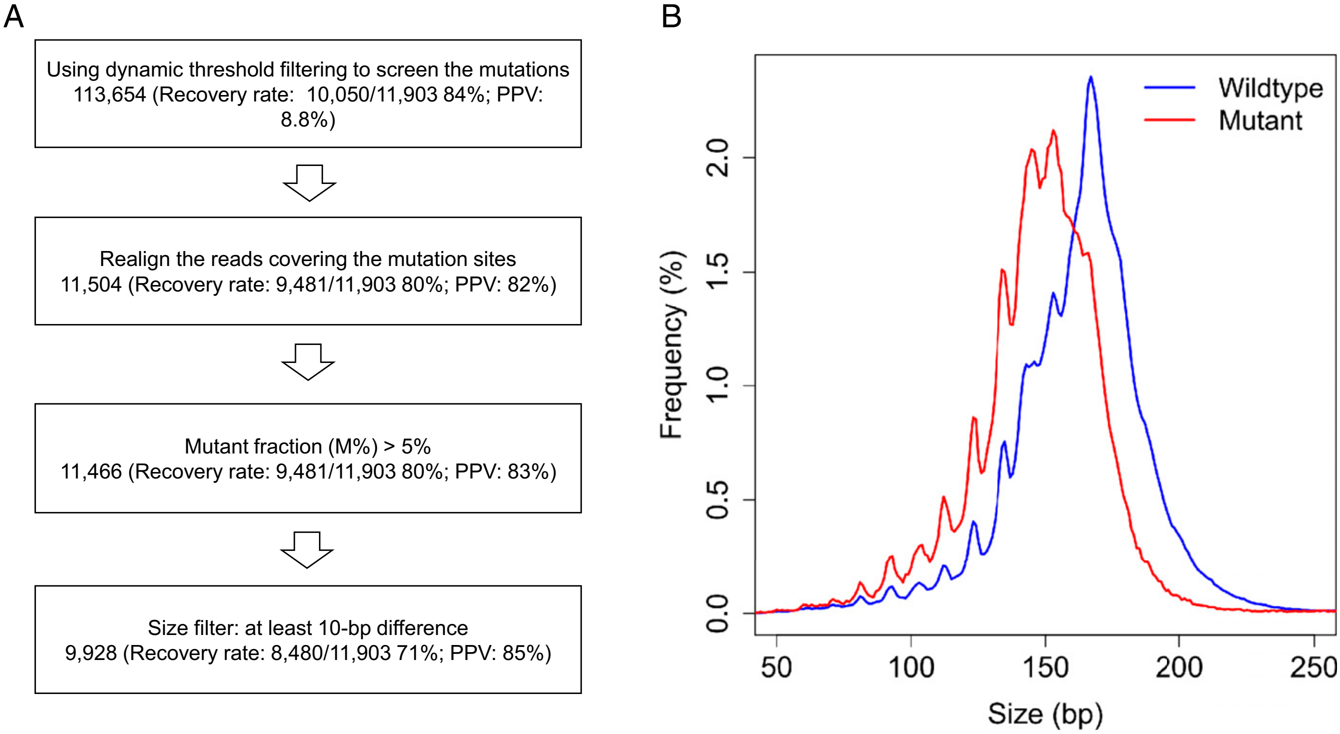

Fig. 1. Identification of somatic mutations in plasma of a HCC patient without a priori tumor information. (*A*) Number of putative variants identified at each successive step of bioinformatics filtering. Recovery rate refers to the proportion of the 11,903 tumor-derived somatic variants that were within the candidate pool of putative variants identified from the cell-free DNA analysis. PPV refers to the number of recovered somatic mutations as a proportion of all putative variants identified. For example, in the last step of size-based filtering, 9,928 putative variants were identified from plasma DNA, among which 8,480 overlapped with the tumor-derived somatic variants (in total 11,903) previously identified from the tumor tissue biopsy. Thus, the recovery rate was 71% (8,480/11,903) and the PPV was 85% (8,480/9,928). (*B*) Frequency distribution of the lengths of cell-free DNA molecules carrying any of the 11,466 putative variants (red curve) identified up to the third filtering step was compared with that of the remaining cell-free DNA molecules (blue curve) between the size ranges of 50 bp and 300 bp.

specificity of the algorithm for identifying tumor-derived somatic variants, we applied the bioinformatics filters to variants observed to be present in the cord plasma but not in the corresponding buffy coat of the cord blood sample. Only 133 variants were finally identified (*SI Appendix*, Fig. S3*A*), and these DNA molecules did not show size shortening (*SI Appendix*, Fig. S3*B*) that was typical of ctDNA. Cell-free DNA of a chronic HBV carrier was also sequenced to 150× haploid genome coverage and analyzed using the same protocol. Only 222 variants were identified, and these DNA molecules did not show size shortening (*SI Appendix*, Fig. S4).

Identification of Liver-Associated Cell-Free DNA End Signatures. Due to the relatively low number of somatic variants associated with a tumor, we looked for other types of ctDNA signatures. Cell-free DNA molecules are short and fragmented, but the fragmentation process is not random (14, 19, 20). We started by exploring the existence of cell-free DNA molecules with liver-associated preferred ends using samples obtained from a liver transplantation donor-recipient pair. Fig. 2*A* is a schematic illustration of the bioinformatics approach. With the use of an Illumina Omni 2.5M single-nucleotide polymorphism (SNP) array, 197,169 SNP loci in which the liver transplant recipient was homozygous (genotype denoted as AA) and the living-related donor was heterozygous (genotype denoted as AB) were identified. A PCR-free plasma DNA library of the liver transplant recipient was sequenced to 260× haploid human genome coverage. Cell-free DNA molecules carrying the B allele were derived from the transplanted liver. Using the B allele fraction, the proportion of liver-derived DNA in the recipient's plasma was 18.3% (21). Among the donor-specific cell-free DNA molecules, we identified end coordinates that were statistically more abundant than predicted by a Poisson probability function if cell-free DNA fragmentation was completely random (14). These sites were termed donor-associated preferred end sites or coordinates. We also identified the recipient-associated preferred end sites using the recipient-specific cell-free DNA molecules that spanned recipient-specific alleles mined from the genotype data.

Fig. 3*A* is a plot of the number of observed cell-free DNA fragments ending at genomic coordinates along a region on chromosome 4. There were loci with overrepresentation of ends of donor-derived DNA as well as loci with overrepresentation of ends of DNA bearing alleles shared by the donor and recipient. We also observed coordinates where there was overrepresentation of cell-free DNA ends irrespective of whether the DNA molecules carried a donor-specific allele or a shared allele. The sizes of cell-free DNA fragments with donor-associated preferred ends were shown to be shorter than those with the recipient-associated preferred ends (Fig. 3*B*), which was consistent with previous data (21).

We identified 6,966 end coordinates (set A) showing statistically significant overrepresentation only among the donor-specific (i.e., liver-specific) cell-free DNA molecules (*SI Appendix*, Fig. S5*A*). Set B included the 21,110 end coordinates that showed statistically significant overrepresentation only among cell-free DNA molecules bearing a shared allele (*SI Appendix*, Fig. S5*A*). 3,525 end coordinates (set C) showed statistically significant overrepresentation among both classes of cell-free DNA molecules, namely those carrying a donor-specific allele and those carrying a shared allele. Because the use of this liver transplantation model to identify cell-free DNA preferred end coordinates was restricted

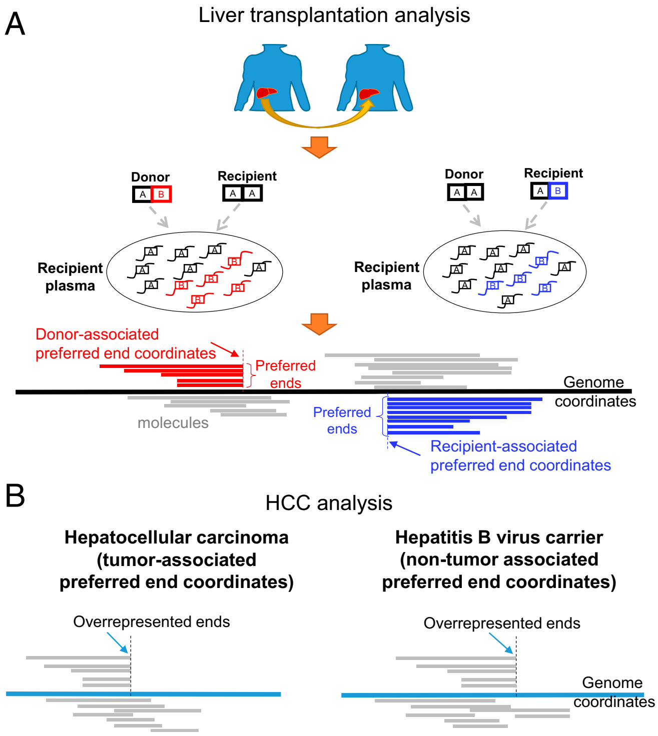
A
Liver transplantation analysis
Donor
A B
Recipient
A A
Donor
A A
Recipient
A B
Recipient plasma
Recipient plasma
Donor-associated preferred end coordinates
Preferred ends
Genome coordinates
molecules
Preferred ends
Recipient-associated preferred end coordinates
B
HCC analysis
Hepatocellular carcinoma (tumor-associated preferred end coordinates)
Hepatitis B virus carrier (non-tumor associated preferred end coordinates)
Overrepresented ends
Overrepresented ends
Genome coordinates

Fig. 2. Schematic illustration of the principle of identifying cell-free DNA end signatures. (*A*) SNP-based end signature analysis. Informative SNP loci where the liver transplant recipient was homozygous (denoted as AA) and the donor was heterozygous (denoted as AB) were used as markers to differentiate and study the donor DNA fragment ends. The B allele (red) represented a donor-specific allele in this context. In the plasma of a liver transplant recipient, the DNA carrying B alleles (red) were donor-derived DNA molecules. The other type of informative SNP loci for which the recipient was heterozygous (denoted as AB) and the donor was homozygous (denoted as AA) were used as markers to differentiate and study the recipient DNA fragment ends. The B allele (blue) represented a recipient-specific allele. In the plasma of a liver transplant recipient, the DNA carrying B alleles (blue) were recipient-derived DNA molecules. Cell-free DNA molecules were aligned and genomic coordinates with overrepresentation, frequency significantly higher than that predicted by the Poisson distribution, of cell-free DNA ends were noted. Genomic coordinates showing a significant overrepresentation of fragment ends associated with donor-derived (red) and recipient-derived (blue) cell-free DNA molecules were termed donor-preferred and recipient-preferred end coordinates, respectively. (*B*) Nonpolymorphism-based end signature analysis. A genomewide scanning strategy was used to identify genomic locations where the observation of cell-free DNA fragment ends were significantly increased compared with that expected for a Poisson distribution in plasma of HCC patients or chronic HBV carriers.

to cell-free DNA molecules carrying a donor-specific allele, we estimated what was the likely proportion of the genome that might bear a liver-associated cell-free DNA preferred end coordinate. The set A preferred end coordinates were associated with 10,438 donor-specific alleles which was 5.3% of the 197,169 donor-specific SNP sites identified from the genomes of this donor-recipient pair. Those 10,438 donor-specific alleles were recovered from sequenced cell-free DNA molecules that spanned 2.3 Mb. Thus, it may be possible that 0.3% (6,966/2.3M) of the coordinates in the genome were liver-associated plasma DNA preferred end coordinates. Multiple cell-free DNA molecules ended on each of the preferred end coordinates. A median of 16 (range 3–82) cell-free DNA molecules terminated at any one of the identified preferred end coordinates. This was 30 times more frequent than if cell-free DNA was randomly cleaved.

We also identified the recipient-associated preferred end sites. Cell-free DNA molecules carrying the recipient-specific alleles were mainly derived from the recipient's hematopoietic system (21). The number of SNPs heterozygous in the recipient (genotype AB) and homozygous in the donor (genotype AA) was 202,652. Set X included 3,036 end coordinates that showed statistically significant overrepresentation among cell-free DNA molecules carrying recipient-specific alleles (*SI Appendix*, Fig. S5*B*). Set Y included 3,445 end coordinates that were significantly overrepresented among cell-free DNA molecules with the shared alleles. Set Z included 2,429 genomic coordinates that were significantly overrepresented among cell-free DNA molecules with a recipient-specific allele as well as those with a shared allele. These preferred end coordinates were obtained from cell-free DNA fragments associated with 2,717 SNPs, accounting for 1.4% of the analyzed SNPs. A median of 18 (range 7–97) cell-free DNA molecules terminated at any one of these preferred end coordinates.

Correlation Between Liver DNA Fraction and Abundance of Cell-Free DNA with Liver-Associated Preferred Ends. The abundance of cell-free DNA molecules with liver-associated preferred ends was calculated based on the number of molecules ending on set A (donor-derived) coordinates as a ratio to the number of molecules ending on set X (recipient-derived) coordinates. We tested if these preferred ends mined from a single individual were also observable in the plasma of other liver transplant recipients. We sequenced the PCR-amplified plasma DNA libraries of 13 other liver transplant recipients (22). The median number of mapped reads was 21.1 million (range: 16.2–26.3 million). Microarray genotyping was also performed for these donor-recipient pairs. The sequence reads aligned to informative SNPs where the donor was heterozygous, and the recipient was homozygous were used to calculate the liver DNA fraction in each plasma sample. A positive correlation was observed between the abundance of cell-free DNA with ends at the liver-associated end coordinates against the liver DNA fractions based on the abundance of donor-specific alleles ($R = 0.78$, $P = 0.0017$, Pearson correlation, Fig. 4). In contrast, *SI Appendix*, Fig. S6 showed that there was no correlation between the SNP-based liver DNA fraction and the abundance of plasma DNA fragments terminating at the fetal-associated preferred end coordinates over the maternal-associated preferred end coordinates identified in a previous study (14).

Tumor-Associated Cell-Free DNA Preferred Ends. Unlike the liver transplantation case where there were 197,169 SNP sites informative for donor-specific alleles, there were just 11,903 somatic mutations harbored by the tumor of the HCC patient. Instead of focusing on cell-free DNA molecules bearing somatic mutations, we compared the ending patterns between cell-free DNA of the HCC patient (220× haploid genome coverage) and cell-free DNA of a patient with chronic HBV infection (150× haploid genome coverage) to identify tumor-associated cell-free DNA molecules (Fig. 2*B*). We chose to compare the HCC cfDNA ending profile against that of a chronic HBV patient because our HCC patient also had chronic HBV infection and we aimed to minimize the chance of labeling HBV-related ending profiles as tumor-associated. We determined if any locations in the human genome showed significantly increased probability beyond that expected by the Poisson distribution of being a cell-free DNA ending site in either sample (15). A P value of <0.01 was used to indicate statistical significance of the overrepresentation of cell-free DNA ends. We found a total of 9.8 and 13.8 million genomic coordinates exhibiting significant overrepresentation of cell-free DNA ends in the HCC sample and HBV sample, respectively. There were 4.4 million preferred end sites common to both samples (Fig. 5*A*). End sites that were not shared between the cell-free DNA samples from the HCC and chronic HBV patients were deemed as the tumor-associated and nontumor-associated preferred end coordinates, respectively. Majority of the tumor-associated preferred end coordinates were found to be in the intergenic regions (58%) or gene bodies (39%) while a small proportion were located at gene promoters (3%).

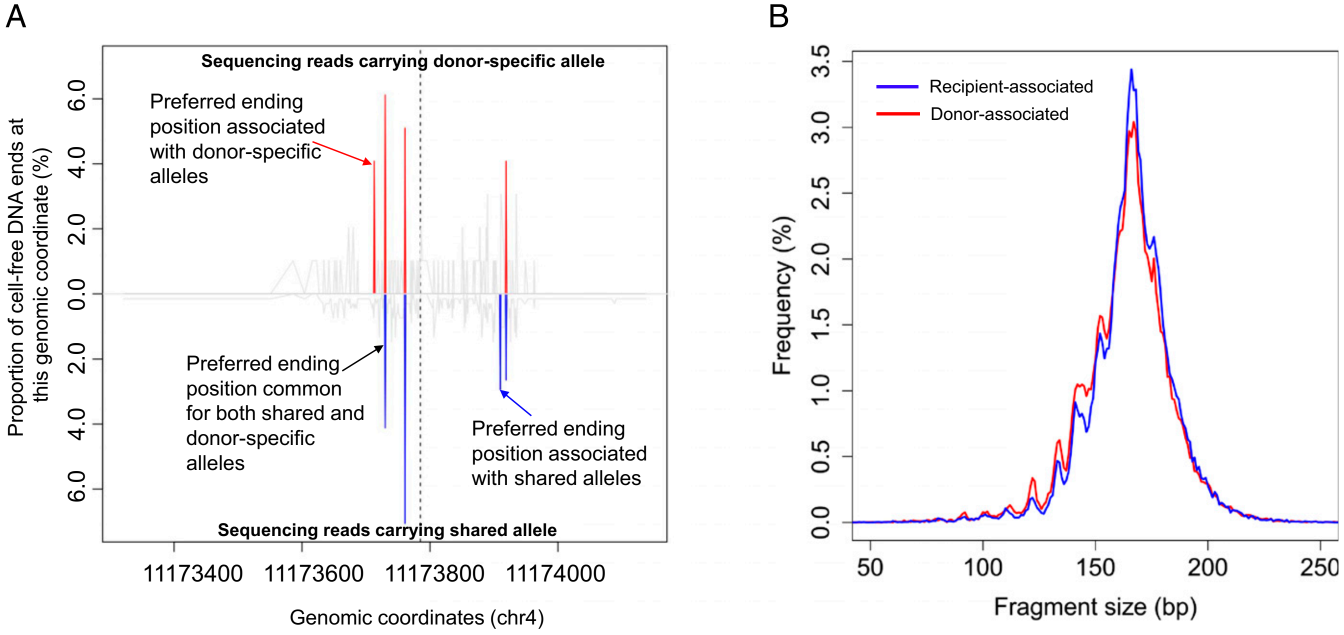

Fig. 3. Cell-free DNA preferred end coordinates identified from the plasma of a liver transplant recipient. (*A*) Plot of frequency of cell-free DNA fragment ends surrounding an informative SNP at which the recipient was homozygous, and the donor was heterozygous along genomic coordinates on chromosome 4. Genomic coordinates with significant overrepresentation of ending positions among cell-free DNA molecules bearing donor-specific alleles were shown in red. Genomic coordinates with significant overrepresentation of ending positions among cell-free DNA molecules bearing shared alleles were shown in blue. Coordinates without overrepresentation of cell-free DNA ends were shown in gray. The position of the SNP of interest was marked by a dotted line. (*B*) Size distributions of cell-free DNA molecules with the donor-associated (red) and recipient-associated preferred end sites (blue).

Evidence for Tumoral Association of the Identified Preferred End Coordinates. Using the preferred end coordinates identified from the discovery sample pair, we calculated the ratios of the number of cell-free DNA molecules showing the tumor-associated preferred ends to those with the nontumor-associated preferred ends among the cell-free DNA samples of 90 HCC patients reported in our previous study (16). Plasma DNA sequencing libraries were prepared using a protocol involving library enrichment by PCR amplification and were sequenced to a median sequencing depth of 1.5× haploid genome coverage (range: 0.8–3.9×). There was a positive correlation between ratios of tumor- to nontumor-associated preferred ends and the tumor DNA fractions ($R = 0.60$, $P < 0.001$, Pearson correlation, Fig. 5*B*). The tumor DNA fractions were determined from quantitative assessment of the extent of chromosomal copy number aberrations as described in our previous study (16). Next, we studied the abundance of the tumor-associated preferred ends and showed that the ratios of tumor- to nontumor-associated preferred ends were significantly increased in the plasma samples of the 90 HCC patients compared with the plasma samples of other non-HCC participants from the same study (16), namely healthy controls ($n = 32$), HBV carriers ($n = 67$), and patients with liver cirrhosis ($n = 36$), (Fig. 6*A*, $P < 0.001$, Kruskal–Wallis test). Using the ratios of the abundance of tumor- to nontumor-associated preferred ends, the area under the receiver operating curve between HCC and non-HCC identification was 0.88 (Fig. 6*B*).

We studied the size profiles of cell-free DNA molecules showing the tumor- and nontumor-associated preferred ends in the previous cohort of 90 cases (16). The size profile of one representative example with plasma tumor DNA fraction of 20.8% is shown in Fig. 7*A*. The size profile for DNA fragments with tumor-associated preferred ends was on the left of that for fragments with nontumor-associated preferred end sites (Fig. 7*A*). These data indicated that cell-free DNA molecules with the tumor-associated preferred ends were shorter than those with the nontumor-associated preferred ends. These results were consistent with the fact that tumor-derived DNA in plasma was shorter than nontumor-derived DNA in plasma (16) and provided additional evidence that the tumor-associated ends that we have identified were likely to be a feature of plasma DNA molecules that originated from the tumor.

To quantify the extent of the size shortening, cumulative frequency plots were generated for the cell-free DNA with the tumor- or nontumor-associated preferred ends (*SI Appendix*, Fig. S7*A*) for

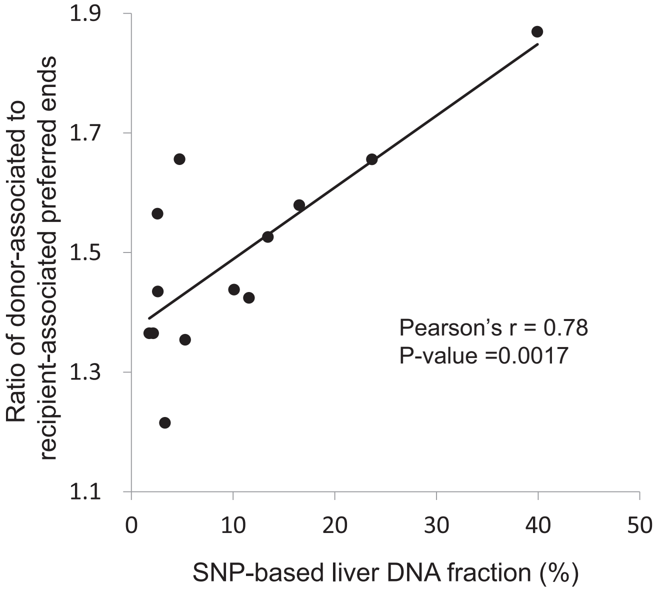

Fig. 4. Correlation analysis between ratios of donor-associated to recipient-associated cell-free DNA preferred ends against SNP-based liver DNA fractions in plasma of liver transplant recipients.

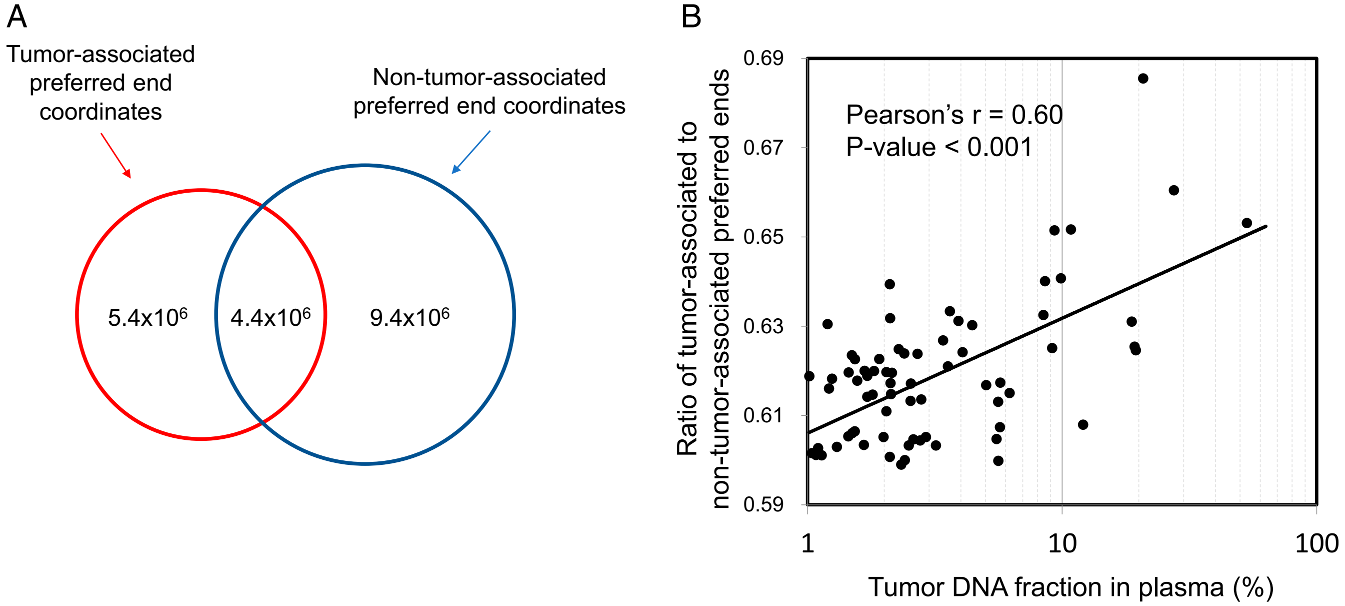

Fig. 5. Tumor-associated and nontumor-associated cell-free DNA preferred end coordinates. (*A*) Number of cell-free DNA preferred end coordinates identified based on comparing the fragment end profiles in plasma of a HCC patient and a chronic HBV carrier. (*B*) Correlation analysis between ratios of tumor- and nontumor-associated cell-free preferred ends against tumor DNA fractions.

each plasma sample. The difference between the two curves, denoted as ΔS (*SI Appendix*, Fig. S7*B*) was calculated (16). A higher positive value of ΔS at a particular size indicated an increased amount of DNA with tumor-associated preferred ends shorter than that size cutoff. The ΔS value at 166 bp, i.e., the ΔS_{166} value, was used to quantify the degree of size shortening among cell-free DNA molecules with tumor-associated preferred ends. The ΔS_{166} values progressively increased as the tumor DNA fraction increased (Fig. 7*B*) in the HCC group. However, for HCC patients with a tumor DNA fraction of less than 2%, there was small reduction in the ΔS_{166} values compared with the non-HCC subjects (Fig. 7*B*). These results were consistent with our previous study whereby plasma from HCC patients with low tumor DNA fraction showed lengthened size profiles, which was speculated to be contributed by cell death in the nonneoplastic liver tissue surrounding the tumor (16).

Discussion

It has been envisioned that ctDNA analysis may provide the means to develop blood-based tests for the early detection of

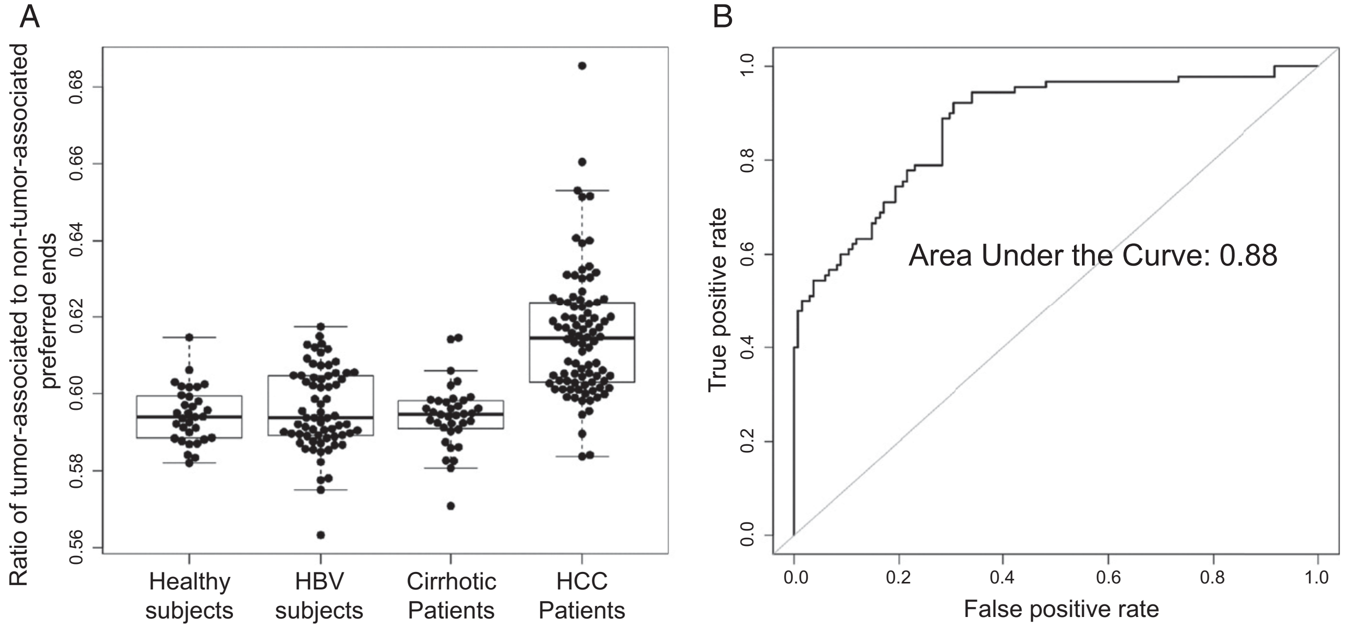

Fig. 6. Abundance of tumor-associated cell-free DNA preferred ends in plasma of HCC and non-HCC patients. (*A*) Ratios of tumor- to nontumor-associated cell-free DNA preferred ends in plasma of healthy subjects, chronic HBV carriers, patients with liver cirrhosis, and HCC patients. (*B*) Receiver-operating curve analysis for discriminating HCC patients from non-HCC subjects using the ratio of tumor- to nontumor-associated cell-free DNA preferred ends.

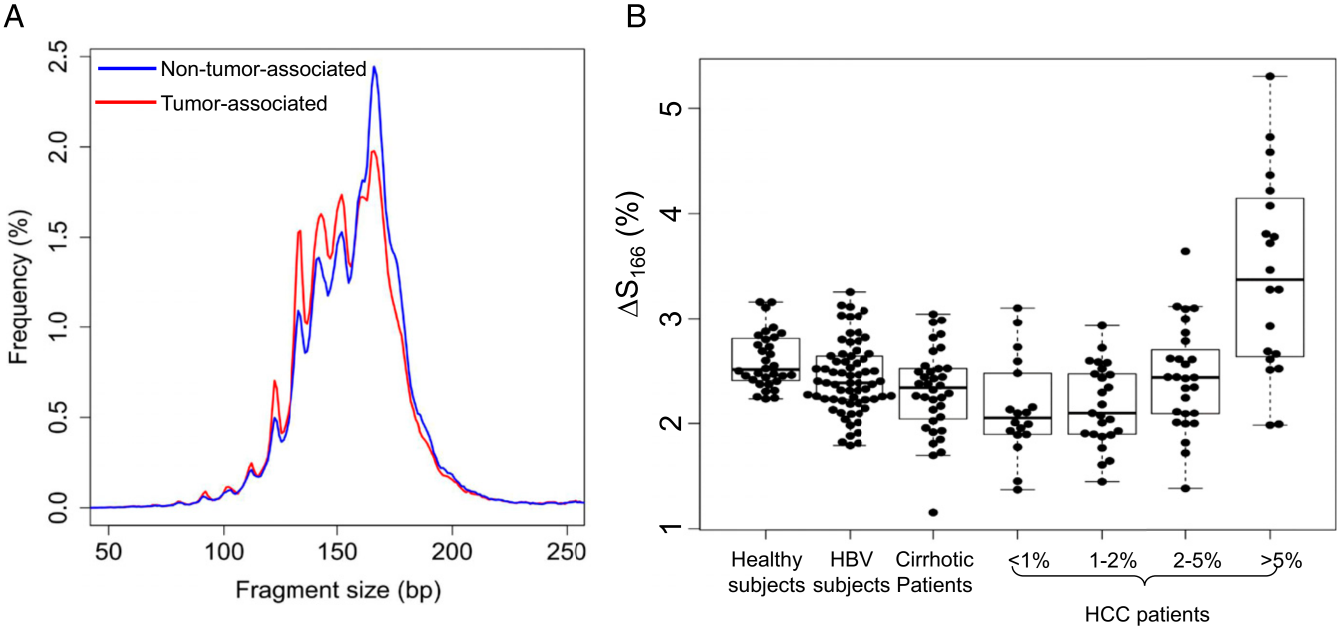

Fig. 7. Size profile analysis of plasma DNA molecules with tumor- or nontumor-associated cell-free DNA preferred ends. (*A*) Size distributions of cell-free DNA with tumor-associated preferred ends (red) and those with nontumor-associated preferred ends (blue). (*B*) ΔS_{166} values for healthy subjects, chronic HBV carriers, patients with liver cirrhosis, and HCC patients.

cancers. However, it is a challenging task to conclusively determine if a nucleic acid signature is a ctDNA molecule when the sample is collected from a person not already known to have cancer. Such a challenge stems from both biological and technical reasons. Biologically, cancers are heterogeneous and their molecular profiles vary from case to case. Besides, if one aims to detect early stage cancers, the tumor might exhibit a smaller subset of the mutations at low concentrations in the circulation (4). Intratumoral heterogeneity may also render the concentrations of ctDNA mutations to be different from each other even in the same plasma sample (7, 8). To enhance the sensitivity of ctDNA analysis for detection of cancers without a priori genetic information from the tumor, we reasoned that one may need to enlarge the pool of cancer-associated qualitative signatures that are searchable in plasma of cancer patients. In this study, we first developed an approach to detect the total pool of somatic mutations, both drivers and passengers, in plasma. Next, we investigated the existence of tumor-associated preferred DNA ends as a class of ctDNA signature.

For somatic mutation detection in plasma, we have adopted a multistep bioinformatics filtering strategy to distinguish true somatic mutations from sequencing artifacts. Each putative variant was compared with the blood cell DNA analysis of the tested individual and would be removed as a candidate somatic mutation if it was present in the blood cell DNA, which may represent a germ-line variant or a variant associated with hematopoietic cell proliferation (23). Because each plasma DNA sample was sequenced to >100× human haploid genome coverage, true somatic mutations should be present among multiple sequence reads covering the locus. However, given the high sequencing depth, the same sequencing error could also be present in more than one read covering the locus due to chance alone. The chance of observing the same sequencing error in multiple reads correlated with the sequencing depth at the base location where the artifactual variant was observed. We have therefore developed a dynamic filtering algorithm that applied a different read threshold depending on the sequencing depth at the base location of each putative somatic variant. A variant would only be retained as a candidate if it was present in multiple reads beyond the threshold for that site. The threshold calculation was based on a generally accepted sequencing error rate of 0.3% for the sequencing platform used in this study.

After the potential sequencing artifacts were removed, a second aligner was used to remove candidate variants that may have evidence of alignment errors. In addition, an abundance filtering was applied to identify variants that were present at no less than a certain fractional concentration. To identify somatic mutations in the tumor biopsy, we adopted an abundance cutoff of 30% because the tumor biopsy should be enriched with tumoral cells, but we also expected there to be intratumoral heterogeneity. For the plasma sample, an abundance cutoff of 5% was used because ctDNA concentrations were more diluted in plasma than tumor tissues. Lastly, because ctDNA molecules were known to be shorter than non-ctDNA molecules in plasma (16), an additional DNA fragment length cutoff could be applied for somatic mutation identification from plasma.

The results showed that this multistep error reduction strategy could substantially enhance the specificity of somatic mutation identification among ctDNA. The PPV improved from 8.8 to 85% (Fig. 1*A*). The candidate somatic variants identified from the tumor tissue or directly from plasma were likely to be true because they were present among cell-free DNA molecules that were shorter than the background DNA molecules (Fig. 1*B* and *SI Appendix*, Fig. S2*A*). Only few false-positive variants were identified using the algorithm in the adjacent normal liver tissue (96 variants), the cord plasma (133 variants), and plasma of a chronic HBV patient (222 variants).

Our strategy was different from methods based on the use of unique molecular barcodes (24). Here, we used PCR-free sequencing libraries to attempt to detect all unique DNA molecules in the sample. On the contrary, methods incorporating the use of molecular barcodes favored amplification of the barcode-tagged DNA molecules (24). Variants observed among sequence reads tagged with the same barcode were deemed as PCR duplicates. However, in our approach, the sequencing libraries were not amplified. Hence, no PCR duplicates would be expected and the sequencing resources could be fully deployed for interrogating the molecular information in the sample.

In fact, the lack of library amplification also facilitated us to identify cell-free DNA end signatures that were overrepresented in plasma. If PCR amplification was adopted, some of such molecules may have been presumed to be PCR duplicates. We embarked on the search for tumor-associated cell-free DNA preferred ends because we reasoned that the total mutational burden of each tumor, reported to be in the order of 3,000–30,000 mutations (12), was too low as a means for cost-effective identification of early cancers through ctDNA analysis. Another class of ctDNA signature was needed to expand the number of tumor-associated hallmarks, which one may be able to detect within a random representative cell-free DNA sample for the identification of early cancers. We explored if tumor-associated cell-free DNA end signatures might exist in the plasma of HCC patients. To test this hypothesis, we first investigated if liver-associated cell-free DNA preferred ends could be detected in plasma of liver transplant recipients. Our data not only showed that they existed, but also that the abundance of cell-free DNA molecules with the liver-associated ends reflected the amount of liver-derived DNA in the plasma samples. Based on those data, we estimated that there could be millions of such liver-associated end coordinates in the genome. Encouraged by those data, we proceeded to identify and demonstrate the presence of tumor-associated cell-free DNA preferred end signatures among HCC patients.

Although we have previously reported the presence of fetus-associated cell-free DNA preferred ends (14), the demonstration of the existence of tumor-associated cell-free DNA preferred ends provided additional biological insights. We showed that the cell-free DNA preferred end coordinates derived from the transplanted liver and the hepatocellular carcinoma were different from those derived from the placenta. From a technical perspective, genotype differences between a fetus and its mother or between a transplant donor and recipient offered a more convenient means to identify cell-free DNA molecules specifically belonging to the relevant individual and, hence, the identification of the preferred end coordinates associated with cell-free DNA released by tissues of that individual. However, a relative disadvantage of the genotype-based approach was that the identification of the preferred end coordinates was limited to cell-free DNA derived from the limited number of informative SNP loci. Consequently, <10,000 preferred end sites were identified to be liver-associated in this study and to be fetus-associated in our previous study (14). It was remarkable to see that subsets of this relatively small set of liver-derived preferred end coordinates were detectable in other unrelated liver transplant donor-recipient pairs (Fig. 4).

Using the number of informative SNPs among the donor-recipient pair and the proportion of such SNPs associated with cell-free DNA preferred end coordinates, we extrapolated the potential number of tissue-specific cell-free DNA preferred end coordinates that might exist in the genome. Based on our data, 0.3% of the genome, meaning about 1 million genomic loci, was expected to be preferred ending locations. Interestingly, we identified 9.8 million and 13.8 million preferred ends in the plasma of the HCC and chronic HBV patients, respectively. It is noteworthy that due to the relatively small number of somatic mutations harbored by the HCC tumor, we did not use the mutations as a means to identify tumor-associated cell-free DNA preferred end coordinates. Instead, a direct comparison between the cell-free DNA profiles between the HCC and chronic HBV patients was performed. This mining strategy was confirmed to be fruitful because the set of tumor-associated preferred end coordinates identified were indeed enriched in the plasma of HCC patients compared with healthy subjects, chronic HBV carriers, and patients with liver cirrhosis (Fig. 6). Quantities of these end signatures bore correlation with the tumor DNA fractions in those samples (Fig. 5*B*). Interestingly, the plasma samples of those 90 HCC patients were only sequenced to 1.5× haploid genome coverage using a library preparation protocol involving PCR amplification. First, these observations supported our hypothesis that having a large pool of candidate ctDNA signatures would facilitate one to gather evidence for the presence of cancer even at a relatively low level of plasma DNA sampling. Second, PCR amplification of the libraries did not hinder the identification of cell-free DNA with those preferred ends. We believe that this was partly because PCR duplicates were not prominent at low sequencing depths and also because the preferred end sites were typically present on just one end of the plasma DNA molecules while PCR duplicates would feature identical sequences on both ends of the cell-free DNA molecules.

We have recently performed a detailed analysis of the plasma DNA preferred ends in maternal plasma (15). We found that fragmentation patterns observed among cell-free fetal DNA and cell-free maternal DNA bore resemblance to the transposase cutting patterns identified from placental tissues and maternal blood cells, respectively. Transposase cuts genomic DNA at locations where the chromatin is accessible. These data suggest that plasma DNA fragmentation may be related to the chromatin accessibility of the tissue of origin. Applying this insight onto the findings of the present study, it is possible that cells of HCC have different chromatin accessibility profiles compared with those of the cells in the transplanted liver. This may further suggest that the profiles of cell-free DNA preferred ends might be different among DNA originating from different organs and tumor derived from different cell types. Thus, the analysis of cell-free DNA preferred ends might enable one to identify the tissue of origin of the associated cell-free DNA molecules and, hence, the organ of pathology or cancer.

In this study, we have demonstrated the presence of tumor-associated cell-free DNA preferred end coordinates among HCC patients and it would be interesting to explore if such observations could be extended to other cancers. In this study, we have explored the potential utilities of detecting such cell-free DNA preferred ends. It would be of value if future studies could be designed to address the clinical value of the detection of cell-free DNA preferred ends in larger sample cohorts. Abundance of these characteristic ends in plasma provided an estimation of the tumor fraction of the sample even at low sequencing depths. This observation provided the evidence to support that these end coordinates could serve as a class of ctDNA signatures, allowing one to determine if a cell-free DNA molecule was likely to be tumor-derived or not. Due to their high abundance across the genome, their sampling and detection could be achieved more cost-effectively than the detection of somatic mutations. One may consider the use of approaches to specifically target the detection of these cancer-associated sequence signatures among cell-free DNA. In fact, many of the strategies formerly developed for the detection of circulating somatic mutations may be adapted for the detection of tumor-associated cell-free DNA preferred ends. With an expanded pool of sequence-specific ctDNA signatures, including somatic mutations and tumor-associated preferred DNA ends, it is expected that the sensitivity of the liquid biopsy approaches could be greatly enhanced, and it would be interesting to assess the value of deploying these end coordinates for the noninvasive detection of early cancers.

Materials and Methods

Sample Processing and Sequencing. Plasma DNA libraries were constructed from 4 mL of plasma without library enrichment, namely without PCR amplification (14). Paired-end massively parallel sequencing was performed. The sequencing data were analyzed using the SOAP2 aligner (17). The paired-end reads were mapped to the reference human genome (hg19) in a paired-end mode, allowing two mismatches for the alignment for each end. Only paired-end reads with both ends aligned to the same chromosome with the correct orientation, spanning an insert size of $\leq$600 bp were used for downstream analysis. Details for the sample processing and sequencing protocols are described in the *SI Appendix, Supplemental Methods*.

Detection of Somatic Mutations. Apparent sequence variants may result from sequencing errors and alignment errors. To remove such apparent variants, we applied a dynamic read count-based filtering strategy followed by sequence read realignment. Variants that were present at lower than a threshold number of reads would be removed from further consideration, whereby the threshold is different depending on the sequence depth at the base position interrogated and, hence, was a dynamic selection of thresholds. The thresholds were selected based on assuming that sequencing errors occurred at a rate of 0.3% (25). Sequence reads containing candidate variants that remained after dynamic filtering were further subjected to realignment using Bowtie2 (18). Sequence reads that were mapped to different genomic locations by the first and second aligners were discarded. Details of the bioinformatics approaches are described in the *SI Appendix, Supplemental Methods*.

Identification of Cell-Free DNA Preferred End Coordinates. In general, the goal of the analysis was to identify genomic coordinates that were statistically significantly overrepresented as a cell-free DNA end position in a sample than as expected by Poisson distribution (14, 15). To identify liver-derived cell-free DNA preferred ends, we compared the cell-free DNA end coordinates among donor-specific and recipient-specific cell-free DNA molecules as determined by SNP genotype differences between the donor and recipient. To study the HCC-derived cell-free DNA preferred ends, we compared the cell-free DNA end coordinates between a plasma sample from an HCC patient and that of a chronic HBV carrier. Details of the bioinformatics approaches are described in *SI Appendix, Supplemental Methods*.

ACKNOWLEDGMENTS. This work was supported by the Research Grants Council of the Hong Kong SAR Government under the Theme-based research scheme T12-401/16-W and a collaborative research grant from Grail.

1. Chan KCA, et al. (2017) Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. *N Engl J Med* 377:513–522.
2. Chan KCA, et al. (2013) Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem* 59:211–224.
3. Chan KCA, et al. (2013) Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci USA* 110:18761–18768.
4. Phallen J, et al. (2017) Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* 9:eaan2415.
5. Heitzer E, Ulz P, Geigl JB, Speicher MR (2016) Non-invasive detection of genome-wide somatic copy number alterations by liquid biopsies. *Mol Oncol* 10:494–502.
6. Bettegowda C, et al. (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 6:224ra24.
7. Newman AM, et al. (2014) An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* 20:548–554.
8. Abbosh C, et al.; TRACERx consortium; PEACE consortium (2017) Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 545:446–451.
9. Lam WKJ, et al. (2018) Sequencing-based counting and size profiling of plasma Epstein-Barr virus DNA enhance population screening of nasopharyngeal carcinoma. *Proc Natl Acad Sci USA* 115:E5115–E5124.
10. Cohen JD, et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359:926–930.
11. Jamal-Hanjani M, et al.; TRACERx Consortium (2017) Tracking the evolution of non-small-cell lung cancer. *N Engl J Med* 376:2109–2121.
12. Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499:214–218.
13. Pleasance ED, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463:191–196.
14. Chan KCA, et al. (2016) Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci USA* 113:E8159–E8168.
15. Sun K, et al. (2018) Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *Proc Natl Acad Sci USA* 115:E5106–E5114.
16. Jiang P, et al. (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci USA* 112:E1317–E1325.
17. Li R, et al. (2009) SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967.
18. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.
19. Underhill HR, et al. (2016) Fragment length of circulating tumor DNA. *PLoS Genet* 12:e1006162.
20. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164:57–68.
21. Zheng YW, et al. (2012) Nonhematopoietically derived DNA is shorter than hematopoietically derived DNA in plasma: A transplantation model. *Clin Chem* 58:549–558.
22. Gai W, et al. (2018) Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. *Clin Chem* 64:1239–1249.
23. Bauml J, Levy B (2018) Clonal hematopoiesis: A new layer in the liquid biopsy story in lung cancer. *Clin Cancer Res* 24:4352–4354.
24. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
25. Ross MG, et al. (2013) Characterizing and measuring bias in sequence data. *Genome Biol* 14:R51.