Clinical Chemistry 59:1
211–224 (2013)

Cancer Diagnostics

fast track

Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing

K.C. Allen Chan,[1,2,3] Peiyong Jiang,[1,2] Yama W.L. Zheng,[1,2] Gary J.W. Liao,[1,2] Hao Sun,[1,2] John Wong,[4] Shing Shun N. Siu,[5] Wing C. Chan,[6] Stephen L. Chan,[3,7] Anthony T.C. Chan,[3,7] Paul B.S. Lai,[4] Rossa W.K. Chiu,[1,2] and Y.M.D. Lo[1,2,3*]

BACKGROUND: Tumor-derived DNA can be found in the plasma of cancer patients. In this study, we explored the use of shotgun massively parallel sequencing (MPS) of plasma DNA from cancer patients to scan a cancer genome noninvasively.

METHODS: Four hepatocellular carcinoma patients and a patient with synchronous breast and ovarian cancers were recruited. DNA was extracted from the tumor tissues, and the preoperative and postoperative plasma samples of these patients were analyzed with shotgun MPS.

RESULTS: We achieved the genomewide profiling of copy number aberrations and point mutations in the plasma of the cancer patients. By detecting and quantifying the genomewide aggregated allelic loss and point mutations, we determined the fractional concentrations of tumor-derived DNA in plasma and correlated these values with tumor size and surgical treatment. We also demonstrated the potential utility of this approach for the analysis of complex oncologic scenarios by studying the patient with 2 synchronous cancers. Through the use of multiregional sequencing of tumoral tissues and shotgun sequencing of plasma DNA, we have shown that plasma DNA sequencing is a valuable approach for studying tumoral heterogeneity.

CONCLUSIONS: Shotgun DNA sequencing of plasma is a potentially powerful tool for cancer detection, monitoring, and research.

The presence of tumor-derived DNA in the plasma of cancer patients offers exciting opportunities for the detection and monitoring of cancer *(1, 2)*. Indeed, cancer-associated microsatellite alterations *(3, 4)*, gene mutations *(5–9)*, DNA-methylation changes *(10, 11)*, and viral nucleic acids *(12)* have been found in the plasma of patients with different cancer types. Most of the previously published work on plasma DNA as a cancer marker has focused on the detection of specific and predetermined molecular targets known to be associated with cancer by means of such methods as the PCR *(3, 4, 12)*, digital PCR *(5–7)*, and digital ligation assays *(9)*. With the advent of massively parallel sequencing (MPS),[8] several groups have incorporated this approach for developing new plasma DNA–based cancer markers. One approach is to use MPS on tumor samples to first identify specific genomic rearrangements that can subsequently be detected in plasma *(13, 14)*. Another approach is based on the use of targeted amplicon sequencing to search for mutations of genes that are commonly found in cancer *(8)*.

[1] Li Ka Shing Institute of Health Sciences, [2] Department of Chemical Pathology, [3] State Key Laboratory in Oncology in South China, Sir Y.K. Pao Centre for Cancer, [4] Department of Surgery, and [5] Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [6] Department of Surgery, North District Hospital, Sheung Shui, New Territories, Hong Kong SAR, China; [7] Department of Clinical Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China.
* Address correspondence to this author at: Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, 30–32 Ngan Shing St., Shatin, New Territories, Hong Kong SAR. Fax +852-26365090; e-mail loym@cuhk.edu.hk.

Received September 6, 2012; accepted September 27, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.196014

[8] Nonstandard abbreviations: MPS, massively parallel sequencing; HCC, hepatocellular carcinoma; SOAP2, Short Oligonucleotide Alignment Program 2; SNP, single-nucleotide polymorphism; LOH, loss of heterozygosity; LOESS, locally weighted scatterplot smoothing; SNV, single nucleotide variant; GAAL, genomewide aggregated allelic loss.

Owing to their targeted nature, the approaches outlined above can provide only a partial glimpse of the tumor genome in the plasma of cancer patients. For a genomewide view of the tumor genome in the circulation, a nontargeted random—or shotgun—sequencing approach would be desirable. In this regard, there has been much progress in the field of noninvasive prenatal diagnosis because of the results obtained with shotgun MPS of DNA from the plasma of pregnant women *(15)*. This approach has allowed the noninvasive detection of fetal chromosomal aneuploidies *(16–18)* and fetal genomic scanning *(19, 20)*.

In this article, we report the use of shotgun MPS to obtain a noninvasive, genomewide view of cancer-associated copy number variations and mutations in DNA in plasma. We have also sought to demonstrate the use of this approach for elucidating important tumoral characteristics, with tumoral heterogeneity used as an example.

Materials and Methods

SAMPLE COLLECTION

Hepatocellular carcinoma (HCC) patients and carriers of chronic hepatitis B were recruited from the Department of Surgery and the Department of Medicine and Therapeutics, respectively, of the Prince of Wales Hospital, Hong Kong, and informed consent and institutional review board approval were obtained. All HCC patients had Barcelona Clinic Liver Cancer stage A1 disease. Informed consent was obtained after the nature and possible consequences of the studies were explained. The patient with synchronous breast and ovarian cancers was recruited from the Department of Clinical Oncology, Prince of Wales Hospital. Peripheral blood samples from all participants were collected into EDTA-containing tubes. The tumor tissues of the HCC patients were obtained during their cancer-resection surgeries.

PROCESSING OF BLOOD

Peripheral blood samples were centrifuged at 1600*g* for 10 min at 4 °C. The plasma portion was recentrifuged at 16 000*g* for 10 min at 4 °C and then stored at −80 °C. Cell-free DNA molecules from 4.8 mL of plasma were extracted according to the blood and body fluid protocol of the QIAamp DSP DNA Blood Mini Kit (Qiagen). The plasma DNA was concentrated with a SpeedVac® Concentrator (Savant DNA120; Thermo Scientific) into a 40-μL final volume per case for subsequent preparation of the DNA-sequencing library.

GENOMIC DNA EXTRACTION

Genomic DNA was extracted from patients' buffy coat samples according to the blood and body fluid protocol of the QIAamp DSP DNA Blood Mini Kit. DNA was extracted from tumor tissues with the QIAamp DNA Mini Kit (Qiagen).

DNA SEQUENCING

Sequencing libraries of the genomic DNA samples were constructed with the Paired-End Sample Preparation Kit (Illumina) according to the manufacturer's instructions. In brief, 1–5 μg genomic DNA was first sheared with a Covaris S220 Focused-ultrasonicator to 200-bp fragments. Afterward, DNA molecules were end-repaired with T_4 DNA polymerase and Klenow polymerase; T_4 polynucleotide kinase was then used to phosphorylate the 5′ ends. A 3′ overhang was created with a 3′-to-5′ exonuclease–deficient Klenow fragment. Illumina adapter oligonucleotides were ligated to the sticky ends. The adapter-ligated DNA was enriched with a 12-cycle PCR. Because the plasma DNA molecules were short fragments *(21)* and the amounts of total DNA in the plasma samples were relatively small, we omitted the fragmentation steps and used a 15-cycle PCR when constructing the DNA libraries from the plasma samples.

An Agilent 2100 Bioanalyzer (Agilent Technologies) was used to check the quality and size of the adapter-ligated DNA libraries. DNA libraries were then measured by a KAPA Library Quantification Kit (Kapa Biosystems) according to the manufacturer's instructions.

The DNA library was diluted and hybridized to the paired-end sequencing flow cells. DNA clusters were generated on a cBot cluster generation system (Illumina) with the TruSeq PE Cluster Generation Kit v2 (Illumina), followed by 51 × 2 cycles or 76 × 2 cycles of sequencing on a HiSeq 2000 system (Illumina) with the TruSeq SBS Kit v2 (Illumina).

SEQUENCE ALIGNMENT AND FILTERING

The paired-end sequencing data were analyzed by means of the Short Oligonucleotide Alignment Program 2 (SOAP2) in the paired-end mode *(22)*. For each paired-end read, 50 bp or 75 bp from each end was aligned to the non–repeat-masked reference human genome (Hg18). Up to 2 nucleotide mismatches were allowed for the alignment of each end. The genomic coordinates of these potential alignments for the 2 ends were then analyzed to determine whether any combination would allow the 2 ends to be aligned to the same chromosome with the correct orientation, spanning an insert size ≤600 bp, and mapping to a single location in the reference human genome. Duplicated reads were defined as paired-end reads in which the insert DNA molecule showed identical start and end locations in the human genome; the duplicate reads were removed as previously described *(19)*.

MICROARRAY ANALYSIS

DNA extracted from the buffy coat and the tumor tissues of the HCC patients was genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0 system, as previously described *(23)*. The microarray data were processed with the Affymetrix Genotyping Console version 4.1. Genotyping analysis and single-nucleotide polymorphism (SNP) calling were performed with the Birdseed v2 algorithm, as previously described *(24)*. The genotyping data for the buffy coat and the tumor tissues were used for identifying loss-of-heterozygosity (LOH) regions and for performing the copy number analysis. Copy number analysis was performed with the Genotyping Console with default parameters from Affymetrix and with a minimum genomic-segment size of 100 bp and a minimum of 5 genetic markers within the segment. Regions with LOH were identified as regions having 1 copy in the tumor tissue and 2 copies in the buffy coat, with the SNPs within these regions being heterozygous in the buffy coat but homozygous in the tumor tissue. For a genomic region exhibiting LOH in a tumor tissue, the SNP alleles that were present in the buffy coat but were absent from or of reduced intensity in the tumor tissues were considered to be the alleles on the deleted segment of the chromosomal region. The alleles that were present in both the buffy coat and the tumor tissue were deemed as having been derived from the nondeleted segment of the chromosomal region.

ARRAY COMPARATIVE GENOMIC HYBRIDIZATION ANALYSIS

DNA samples extracted from the buffy coat and the tumor tissues of the HCC patients were analyzed with the SurePrint G3 Human High Resolution Microarray Kit (Agilent) as previously described *(25)*. Array comparative genomic hybridization data for the HCC patients were analyzed for copy number variation with the Partek® Genomics Suite. In brief, the raw probe intensities were adjusted according to the GC content of the sequence. This adjustment was followed by probe-level normalization of signal intensity while simultaneously adjusting for fragment length and probe sequences across all samples. Copy number gains and losses were detected by applying the default parameters of the Genomic Segmentation algorithm available in Partek Genomics Suite version 6.5 to obtain the different partitions of the copy number state.

DETECTION OF COPY NUMBER ABERRATION IN TUMOR TISSUE SAMPLES BY SEQUENCING

To investigate genomic copy number aberrations (e.g., copy number gains and copy number losses), we divided the genome into equal-sized segments (1 Mb per window/bin), and tallied the numbers of sequence reads mapping to each bin. Owing to the presence of GC-dependent sequencing biases with high-throughput sequencing technologies *(26)*, we used a statistical correction method, locally weighted scatterplot smoothing (LOESS), to correct the GC-associated bias *(27)*. In this method, a correction factor is calculated for each bin according to the LOESS regression model, as previously described *(28)*. Then, the read counts of each bin were adjusted with the bin-specific correction factor and normalized with the median read counts of all bins. After GC correction, a ratio of the adjusted read counts of the tumor to those of the buffy coat was calculated with the following equation:

$$R = \frac{A_{\text{tumor}}}{A_{\text{BC}}},$$

where A_{tumor} is the normalized GC-adjusted read counts of the tumor tissue and A_{BC} is the normalized GC-adjusted read counts of the buffy coat.

We then constructed a frequency distribution of $\log_2(R)$ for all bins. This distribution plot was used to estimate the proportion of tumor cells (F) in the tumor tissue that showed a particular copy number distribution and, subsequently, the copy number change at each bin.

On the frequency-distribution plot, a central peak at which R is approximately equal to 1 [i.e., $\log_2(R) = 0$] was identified; this peak represents the genomic regions without copy number aberrations. Then, the peaks lying to the left and right of the central peak were identified. These peaks represented regions with a 1-copy loss and a 1-copy gain, respectively. The distances of the left and right peaks from the central peak were used to determine the proportion of tumor cells (F) in the tumor tissue, according to the following equation:

$$F = R_{\text{right}} - R_{\text{left}},$$

where R_{right} is the R value of the right peak and R_{left} is the R value of the left peak.

Then, the copy number change (CN) values for all 1-Mb bins were calculated with the following equation:

$$\text{CN} = \frac{R_{\text{bin}} - R_{\text{cen}}}{0.5 \times F},$$

where R_{bin} is the R value of the bin and R_{cen} is the R value of the central peak.

DETECTION OF COPY NUMBER ABERRATIONS IN PLASMA

We analyzed the genomic representation of plasma DNA for different genomic regions. First, the entire genome was divided into 1-Mb windows, similar to the analysis of copy number aberrations in the tumor tissues. The GC-corrected read count was then determined (as described above) for each 1-Mb window. A z

score statistic was used to determine if the plasma DNA representation in a 1-Mb window would be significantly increased or decreased when compared with the reference group. The reference group consisted of the plasma samples from 16 healthy control individuals. In the current study, the GC-corrected read counts of each 1-Mb bin were normalized to the median GC-corrected read counts of all bins in the sample. The normalized plasma DNA representation was then compared with the data from the controls. A z score was then calculated for each 1-Mb window by using the mean and SDs of the controls. Regions with z scores of <-3 and >3 were regarded as significantly under- and overrepresented, respectively.

NUMBER OF MOLECULES REQUIRED FOR IDENTIFYING COPY NUMBER ABERRATIONS IN PLASMA

For copy number aberration analysis, the sensitivity and specificity of detecting tumor-associated copy number aberrations in plasma were determined by the precision of measuring the representation of plasma DNA in a chromosomal region and the fractional concentration of the tumor-derived DNA in the plasma of the cancer patient. The precision of measuring the plasma DNA representation in turn was affected by the number of plasma DNA molecules analyzed. In this regard, we performed simulation analyses to determine the relationship between the number of plasma DNA molecules required for analysis and the fractional concentration of tumor-derived DNA in the plasma so we could achieve a sensitivity of 95% for the detection of tumor-associated copy number aberrations. Computer simulations were performed for scenarios in which the affected region had a copy number change of -1, $+1$, and $+2$ and for fractional concentrations of tumor-derived DNA ranging from 1% to 50%. In each simulation analysis, the entire genome was divided into 3000 bins. This number was similar to the one we used in the actual experimental analysis when a 1-Mb resolution was used.

We assumed that 10% of the bins would exhibit chromosomal aberrations in the tumor tissue. In the tumor tissue, the expected fraction (P) of total molecules falling into a bin within an affected region would be:

$$P = \frac{2 + \mathrm{CN}}{2 \times 3000 + 3000 \times 10\% \times \mathrm{CN}},$$

where CN is the copy number change. From this information, we calculated the expected change in the plasma.

In the plasma, the expected proportion of the total molecules (E) falling into a bin within an affected region can be calculated as:

$$E = P \times f + \frac{1 - f}{3000},$$

where f is the fractional concentration of tumor-derived DNA in plasma.

Simulations of 1000 normal cases and 1000 cancer cases were performed on the assumption of a binomial distribution of the plasma DNA molecules, with the expected plasma representations as calculated above and with an increasing number of molecules being analyzed until the 95% detection rate was reached. The simulation was conducted with the rbinom function in R (http://www.r-project.org/).

DETECTION OF TUMOR-ASSOCIATED SINGLE-NUCLEOTIDE VARIANTS

We sequenced the paired tumor and constitutional DNA samples to identify the tumor-associated single-nucleotide variants (SNVs). We focused on the SNVs occurring at homozygous sites in the constitutional DNA (i.e., buffy coat DNA). In principle, any nucleotide variation detected in the sequencing data of the tumor tissues but absent in the constitutional DNA could be a potential mutation (i.e., a SNV). Because of sequencing errors (0.1%–0.3% of sequenced nucleotides) *(29)*, however, millions of false positives would be identified in the genome if a single occurrence of any nucleotide change in the sequencing data of the tumor tissue were to be regarded as a tumor-associated SNV. One way to reduce the number of false positives would be to institute the criterion of observing multiple occurrences of the same nucleotide change in the sequencing data in the tumor tissue before a tumor-associated SNV would be called. Because the occurrence of sequencing errors is a stochastic process, the number of false positives due to sequencing errors would decrease exponentially with the increasing number of occurrences required for an observed SNV to be qualified as a tumor-associated SNV. On the other hand, the number of false positives would increase exponentially with increasing sequencing depth. These relationships could be predicted with Poisson and binomial distribution functions. In this regard, we have developed a mathematical algorithm to determine the dynamic threshold of occurrence for qualifying an observed SNV as tumor associated. This algorithm takes into account the actual coverage of the particular nucleotide in the tumor sequencing data, the sequencing error rate, the maximum false-positive rate allowed, and the desired sensitivity for mutation detection.

In this study, we set very stringent criteria to reduce false positives. We required a mutation to be completely absent in the constitutional DNA sequencing, and the sequencing depth for the particular nucleotide position had to be >20-fold. This threshold of occur-

rence was required to control the false-positive detection rate at $<1 \times 10^{-7}$. In this algorithm we also filtered out SNVs that were within centromeric, telomeric, and low-complexity regions to minimize false positives due to alignment artifacts. In addition, putative SNVs mapping to known SNPs in the dbSNP build 135 database were also removed.

Results

TUMOR-ASSOCIATED COPY NUMBER ABERRATIONS IN PLASMA

We investigated whether tumor-associated copy number aberrations could be detected in the plasma of cancer patients by shotgun MPS. Peripheral blood samples were obtained both before and 1 week after surgical resection with curative intent from 4 HCC patients. The blood samples were fractionated into plasma and blood cells. DNA was also obtained from each of the tumors. Copy number aberrations in the 4 tumor samples were analyzed with MPS and with 1 or 2 microarray platforms (Affymetrix and Agilent). Copy number aberrations were analyzed in 1-Mb windows across the genome in the tumor tissues and compared with the plasma samples from a group of 16 healthy control individuals. The data were consistent across the 3 platforms (see Fig. 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol59/issue1).

We then used MPS to analyze the pre- and post-resection plasma samples obtained from all 4 HCC patients. The mean sequencing depth was 17-fold coverage of the haploid human genome (range, 15.2-fold to 18.5-fold). Fig. 1 shows Circos plots *(30)* of the copy number aberrations across the genome in the tumor, the pre-resection plasma sample, and the post-resection plasma sample, for each patient. In each case, characteristic copy number aberrations seen in the tumor tissue sample were also observed in the pre-resection plasma sample (Fig. 1). A significant change in the regional representation of plasma DNA was defined as >3 SDs from the mean representation of the 16 healthy controls for the corresponding 1-Mb window.

For all cases, such copy number aberrations disappeared almost completely in the post-resection plasma sample (Fig. 1). The detectability of the different classes of tumor-associated genetic alterations in plasma is shown in Fig. 2. For comparison, we used the same approach to analyze plasma DNA samples from 4 hepatitis B carriers without HCC (Fig. 1E; see Fig. 2 in the online Data Supplement). These individuals were followed up for 1 additional year after blood sampling and had no evidence of HCC. For these individuals, 99% of the sequenced bins showed normal representations in plasma (see Table 1 in the online Data Supplement). Similarly, a mean of 98.9% of the sequenced bins in the 16 healthy controls showed normal representations in plasma (see Table 2 and Fig. 3 in the online Data Supplement). These results indicate that the analysis of copy number aberrations in plasma is specific for differentiating between cancer patients and individuals without a cancer; however, the specificity for plasma copy number analysis appeared to be reduced in the HCC patients. Hence, in the 4 HCC patients, a median of 15% (range, 2%–48%) of the regions at which no copy number aberrations occurred in the corresponding tumor tissue showed an aberrant plasma DNA representation (see Table 1 in the online Data Supplement). This issue will be discussed in more detail in the Discussion section.

FRACTIONAL CONCENTRATION OF TUMOR DNA IN PLASMA DETERMINED BY GENOMEWIDE AGGREGATED ALLELIC LOSS ANALYSIS

The fractional concentrations of tumor-derived DNA in plasma were determined by analyzing, in a genomewide manner, the allelic counts for SNPs exhibiting LOH in the plasma shotgun MPS data, which we term "genomewide aggregated allelic loss" (GAAL) analysis. For such an analysis, we chose SNPs that exhibited LOH in the tumors as demonstrated with the Affymetrix SNP 6.0 microarray. The alleles deleted in the tumors would have lower concentrations in the plasma than those that were not deleted. The difference in their concentrations was related to the concentration of tumor-derived DNA in the plasma sample. Thus, the plasma concentration of the tumor-derived DNA (C) can be deduced with the following equation:

$$C = \frac{N_{\text{nondel}} - N_{\text{del}}}{N_{\text{nondel}}},$$

where N_{nondel} represents the number of sequenced reads carrying the nondeleted alleles in the tumor tissues, and N_{del} represents the number of sequenced reads carrying the deleted alleles in the tumor tissues.

Table 1 lists the fractional concentrations of tumor DNA in the plasma samples for each of the 4 cases. The size of the tumor appears to be correlated with the estimated fractional concentration of tumor-derived DNA in plasma before surgical resection. For example, we estimated that tumor-derived DNA accounted for 52% of the total plasma DNA in the patient who had the largest tumor (13 cm) of the 4 HCC cases. For each of the 4 cases, we observed a reduction in the fractional concentration of tumor-derived DNA after surgical resection of the tumor (Table 1).

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.