RESEARCH ARTICLE

DIAGNOSTICS

# Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing

Rebecca J. Leary,[1]* Mark Sausen,[1]* Isaac Kinde,[1]* Nickolas Papadopoulos,[1] John D. Carpten,[2] David Craig,[2] Joyce O'Shaughnessy,[3] Kenneth W. Kinzler,[1] Giovanni Parmigiani,[4,5] Bert Vogelstein,[1] Luis A. Diaz Jr.,[1†] Victor E. Velculescu[1†]

**Clinical management of cancer patients could be improved through the development of noninvasive approaches for the detection of incipient, residual, and recurrent tumors. We describe an approach to directly identify tumor-derived chromosomal alterations through analysis of circulating cell-free DNA from cancer patients. Whole-genome analyses of DNA from the plasma of 10 colorectal and breast cancer patients and 10 healthy individuals with massively parallel sequencing identified, in all patients, structural alterations that were not present in plasma DNA from healthy subjects. Detected alterations comprised chromosomal copy number changes and rearrangements, including amplification of cancer driver genes such as *ERBB2* and *CDK6*. The level of circulating tumor DNA in the cancer patients ranged from 1.4 to 47.9%. The sensitivity and specificity of this approach are dependent on the amount of sequence data obtained and are derived from the fact that most cancers harbor multiple chromosomal alterations, each of which is unlikely to be present in normal cells. Given that chromosomal abnormalities are present in nearly all human cancers, this approach represents a useful method for the noninvasive detection of human tumors that is not dependent on the availability of tumor biopsies.**

## INTRODUCTION

Abnormal chromosome content, or aneuploidy, is a common characteristic of tumors, which manifests at the earliest stages of tumorigenesis and increases throughout subsequent tumor development (*1*–*4*). In addition to losses and gains of entire chromosomes, alterations of chromosome arms, focal amplifications and deletions, and rearrangements are observed in nearly all cancer genomes. Analysis of such alterations in cancer began with karyotyping but is now generally carried out with molecular methods that can more easily assess genomes in a comprehensive manner. For example, an approach based on sequencing and enumerating genomic DNA tags, called digital karyotyping (DK), was developed for the analysis of copy number alterations on a genome-wide scale (*5*). Similar tag-based approaches have been adapted to next-generation sequencing methods (*6*, *7*). Likewise, the analysis of chromosomal rearrangements with large-scale DNA sequencing approaches allows for high-resolution mapping of rearrangement breakpoints (*3*).

Given the universal nature of chromosomal alterations in human cancer and improved methods for detecting such changes, we wondered whether we could directly identify chromosomal alterations in the circulation of cancer patients. Sequencing analyses of chromosome content in the maternal circulation are now being used for detection of fetal aneuploidy (*8*, *9*), although such approaches have not been evaluated for detection of chromosomal alterations in cancer patients. Similarly, analysis of circulating tumor DNA in patients with hematopoietic malignancies has been useful for the detection of known recurrent chromosomal rearrangements, such as those that involve the *BCR-ABL* oncogene and genes that encode immunoglobulin chains, T cell receptor subunits, and the retinoic acid receptor (*10*–*15*). More recently, analysis of tumor rearrangements has allowed the development of patient-specific biomarkers that can be evaluated in plasma for the detection of residual disease or for tumor monitoring (*6*, *16*). However, such monitoring approaches rely on analyses of known alterations identified in resected tumors from the same patients and cannot be directly applied to the detection of new alterations in the circulation of patients in whom biopsied material is unavailable. Recurrent mutations, including those identified in oncogenes such as *KRAS*, have also been readily identified in a fraction of patients with solid tumors (*17*–*19*).

An alternative to these approaches is the identification of de novo tumor-derived chromosomal alterations through massively parallel direct sequencing of DNA from the circulation of cancer patients. Such approaches would be applicable to more patients than those that rely on recurrent oncogene alterations and could theoretically permit noninvasive detection of nearly all cancer types. Herein, we compare whole-genome analyses of DNA from the plasma of late-stage cancer patients to healthy individuals with massively parallel sequencing and detect structural alterations specific to patients.

[1]Ludwig Center for Cancer Genetics and Howard Hughes Medical Institutions, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21287, USA. [2]Translational Genomics Research Institute, Phoenix, AZ 85044, USA. [3]Baylor Sammons Cancer Center, Texas Oncology, US Oncology, Dallas, TX 75246, USA. [4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [5]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.
*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: velculescu@jhmi.edu (V.E.V.); ldiaz1@jhmi.edu (L.A.D.)

## RESULTS

### Overview

A schematic of our approach to examine chromosomal abnormalities directly in the plasma of cancer patients is illustrated in Fig. 1. As a

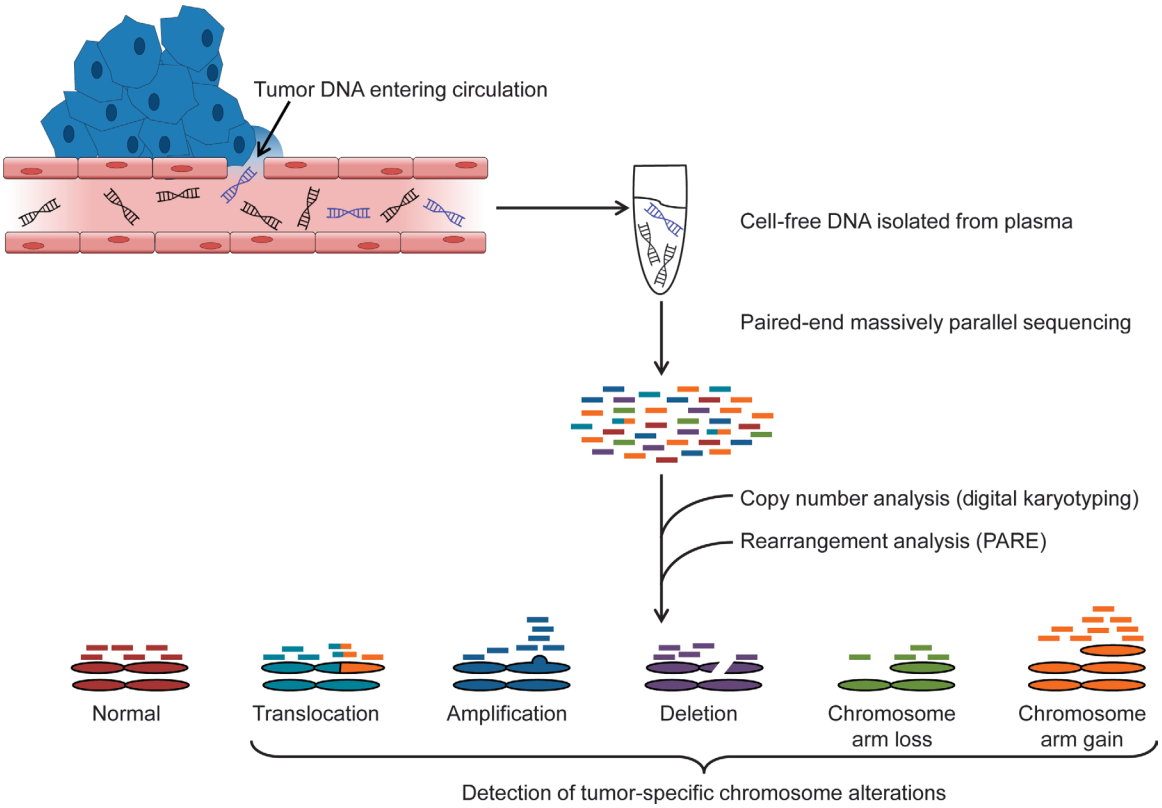



**Fig. 1.** Schematic of analyses for direct detection of chromosomal alterations in plasma. The method uses next-generation paired-end sequencing of cell-free DNA isolated from plasma to identify chromosomal alterations characteristic of tumor DNA. Such alterations include copy number changes (gains and losses of chromosome arms) as well as rearrangements resulting from translocations, amplifications, or deletions.

proof-of-principle analysis, we obtained 4 to 18 ml of plasma from each of 10 healthy individuals (N1 to N10), 7 patients with colorectal cancer (CRC11 to CRC17), and 3 patients with breast cancer (BR1 to BR3) (table S1). Plasma DNA was purified and used to generate paired-end libraries for whole-genome sequencing, and each library was analyzed on two lanes of an Illumina HiSeq instrument (see Materials and Methods). An average of 249,378,422 distinct paired sequences [50 base pairs (bp) from each end] was obtained for each sample (Table 1). The resulting sequence data from circulating DNA were analyzed for chromosome copy number changes and for intra- and interchromosomal rearrangements.

### Analysis of chromosomal copy number changes

Losses or gains of specific chromosomal regions are a hallmark of many cancers and have been used historically to identify tumor suppressors or oncogenes targeted by the alterations (*20*–*22*). Such chromosomal imbalances could be useful as markers of tumorigenesis because they should, in principle, alter the chromosomal representation of circulating DNA. To adapt DK to detect tumor-specific (somatic) chromosomal alterations in the plasma, we used the equivalent of one lane of HiSeq single-read sequence data per sample (average of 144,543,191 distinct reads) and applied a number of filtering steps to remove sources of variation that were not tumor-specific (see Materials and Methods). For example, we removed sequences that are known to vary in the germlines of normal individuals, because these could confound identification of somatic copy number changes. In addition, we applied a weight to each sequence read based on local GC content. This weighting has been shown to remove bias introduced by next-generation sequencing and allows for a more accurate assessment of chromosomal representation of the original genomic DNA (see Materials and Methods) (*23*, *24*). The resulting weighted reads were used to determine the proportion of reads that mapped to specific regions in the genome (fig. S1). We performed analyses of entire chromosomes, of chromosome arms, and of sequential regions of specified sizes (for example, 10 Mb) throughout the genome. Although each of these approaches has certain advantages, we chose to analyze chromosome arms because these were frequently altered in breast and colorectal cancer samples previously analyzed for copy number alterations and would be expected to be altered in most human cancers (see Materials and Methods).

The proportion of sequences that represented each chromosome arm (excluding short arms of acrocentric chromosomes) was calculated, for each sample, by dividing the sum of the weighted reads mapping to that arm by the total number of weighted reads mapping to the reference genome. For the normal samples, N1 to N10, the proportion of chromosomal arm sequences ranged from 0.46 to 6.19%, closely corresponding to the expected fraction based on genomic size and the applied mapping criteria (table S2) ($R^2 = 0.95$; $P < 0.0001$, Pearson correlation). The variation among the normalized proportions of each chromosomal arm in the plasma from normal individuals was very low (average, 2.56 ± 0.0065%; range of SD, ±0.0025% to ±0.014%). These results are consistent with similar measurements of

**Table 1.** Summary of next-generation sequencing analyses performed. Data were obtained using next-generation sequencing analyses performed on Illumina HiSeq instruments using 50-bp PE reads. Distinct paired reads correspond to read pairs having unique genomic start sites. Sequence coverage indicates average number of reads per base per haploid genome. Physical coverage indicates average number of paired reads spanning any base in a haploid genome assuming a 165-bp fragment size.

| Sample name | Patient diagnosis | Sample origin | Total bases sequenced | Total distinct paired reads | Sequence coverage | Physical coverage |
|---|---|---|---|---|---|---|
| CRC11 | Colorectal cancer | Plasma | 24,728,144,682 | 216,092,204 | 8.2 | 11.9 |
| CRC12 | Colorectal cancer | Plasma | 25,707,029,400 | 221,881,499 | 8.6 | 12.2 |
| CRC12-PT | Colorectal cancer | Tumor | 14,984,097,228 | 113,414,859 | 5.0 | 6.2 |
| CRC13 | Colorectal cancer | Plasma | 24,033,905,652 | 206,926,843 | 8.0 | 11.4 |
| CRC14 | Colorectal cancer | Plasma | 23,774,411,124 | 201,571,426 | 7.9 | 11.1 |
| CRC14-0 | Colorectal cancer | Plasma | 3,113,755,960 | 17,297,631 | 1.0 | 1.0 |
| CRC14-4 | Colorectal cancer | Plasma | 3,755,921,750 | 21,193,176 | 1.3 | 1.2 |
| CRC14-PT | Colorectal cancer | Tumor | 7,156,542,105 | 55,882,387 | 2.4 | 3.1 |
| CRC15 | Colorectal cancer | Plasma | 34,216,224,513 | 299,895,779 | 11.4 | 16.5 |
| CRC15-PT | Colorectal cancer | Tumor | 12,466,375,200 | 102,193,630 | 4.2 | 5.6 |
| CRC16 | Colorectal cancer | Plasma | 32,670,584,037 | 284,131,684 | 10.9 | 15.6 |
| CRC16-PT | Colorectal cancer | Tumor | 6,199,959,150 | 41,700,493 | 2.1 | 2.3 |
| CRC17 | Colorectal cancer | Plasma | 33,060,006,522 | 286,524,067 | 11.0 | 15.8 |
| BR1 | Breast cancer | Plasma | 34,918,959,429 | 304,798,840 | 11.6 | 16.8 |
| BR2 | Breast cancer | Plasma | 30,171,911,865 | 240,263,505 | 10.1 | 13.2 |
| BR3 | Breast cancer | Plasma | 31,294,294,671 | 259,443,659 | 10.4 | 14.3 |
| N1 | Normal | Plasma | 26,918,560,359 | 231,520,314 | 9.0 | 12.7 |
| N2 | Normal | Plasma | 25,928,759,499 | 224,708,017 | 8.6 | 12.4 |
| N3 | Normal | Plasma | 21,331,401,576 | 161,874,934 | 7.1 | 8.9 |
| N4 | Normal | Plasma | 25,735,971,696 | 223,342,309 | 8.6 | 12.3 |
| N5 | Normal | Plasma | 33,535,967,796 | 288,318,721 | 11.2 | 15.9 |
| N6 | Normal | Plasma | 32,892,667,872 | 285,785,107 | 11.0 | 15.7 |
| N7 | Normal | Plasma | 27,558,615,816 | 226,653,076 | 9.2 | 12.5 |
| N8 | Normal | Plasma | 32,472,392,886 | 279,224,469 | 10.8 | 15.4 |
| N9 | Normal | Plasma | 30,058,183,548 | 257,431,312 | 10.0 | 14.2 |
| N10 | Normal | Plasma | 33,068,060,850 | 287,180,675 | 11.0 | 15.8 |

circulating DNA from the plasma of pregnant women carrying euploid fetuses (*8*, *9*). In contrast, the normalized proportions of chromosomal arm sequences in the plasma of cancer patients were much more variable, ranging from 0.61- to 1.97-fold of the average found in the plasma of normal individuals (table S2).

To determine whether sequenced reads for an individual patient sample deviate from patterns in normal samples, we used the fraction of reads that mapped to each arm to calculate a $z$ score. For each arm, the $z$ score was calculated as the number of SDs from the mean of the reference plasma samples (N1 to N10). After Bonferroni correction for multiple comparisons of the 39 chromosomal arms, an absolute $z$ score of $\geq$11.88 was determined to represent a statistically significant gain or loss of a chromosomal arm ($P < 0.05$, Student's $t$ test). All chromosome arms of the 10 normal plasma samples had absolute $z$ scores of less than 2.62. In contrast, plasma samples from all 10 of the cancer patients showed evidence of copy number gains or losses, with the highest absolute $z$ score in each sample ranging from 13.3 to 434.4 (Fig. 2A).

Although such analyses could be used to evaluate specific chromosomal arms, a statistical approach that uses a combination of the most markedly altered chromosome arms in each sample would be expected to provide a more sensitive measure of circulating tumor DNA. We analyzed previously obtained genome-wide copy number alterations detected from single-nucleotide polymorphism (SNP) arrays of 36 colorectal cancer samples (*25*) to determine how frequently tumors lost multiple chromosome arms. As shown in fig. S2, we found that the mean number of chromosome arms altered in these colorectal cancers was 21 and ranged from 5 to 35. Accordingly, we constructed a log-scale plasma aneuploidy score (PA score) based on the five chromosomes whose arms had the highest absolute $z$ scores (see Materials and Methods). The PA score from the plasma of healthy individuals ranged from 0.1 to 2.4, and we calculated that a threshold PA score of

5.84 would provide a specificity of >99% (Student's *t* distribution) for indicating aneuploidy (Fig. 2B). All plasma samples from the colorectal and breast cancer patients had PA scores above this threshold, ranging from 11.9 to 41.5 (Fig. 2B and tables S1 and S2). The two plasma samples with the lowest PA scores represented those with the lowest amounts of circulating tumor DNA, and the PA score generally correlated with tumor burden ($R^2 = 0.53$; $P = 0.017$, Pearson correlation) (Fig. 2B, table S2, and Materials and Methods).

### Analysis of rearrangements

The chromosomal instability that underlies large chromosomal gains and losses in tumorigenesis is associated with genomic rearrangements. Such somatic rearrangements are not present in normal cells in a clonal fashion and would therefore be expected to provide a highly sensitive and specific marker for the presence of clonal tumor-specific genetic alterations. We previously developed a technique, personalized analysis of rearranged ends (PARE), to identify rearranged breakpoints from tumor DNA for individual patients (see Materials and Methods). A challenge in adapting PARE to detection of rearrangements directly from plasma DNA is distinguishing the relatively few somatic rearrangements present in circulating tumor DNA from the much larger number of structural variants resulting from copy number variations in the germline of all individuals. To overcome this obstacle, we used bioinformatic filters that enriched for high-confidence somatic structural alterations while removing germline and artifactual changes. These filters included selecting paired-end reads that (i) mapped to different chromosomes or to the same chromosome but at large distances (≥30 kb) apart, (ii) spanned rearrangement junctions that were observed in multiple reads, (iii) contained sequenced rearrangement breakpoints, and (iv) mapped to genomic regions that did not contain known germline copy number variants or repeated sequences (*26*, *27*) (fig. S1).

Paired-end Illumina sequence data for DNA in plasma samples from the 10 cancer patients and 10 healthy individuals (table S1) revealed a total of 65,402,563 aberrantly mapped paired-end reads, most of which were expected to result from either germline changes or mapping artifacts (*26*, *27*). Application of the criteria described above identified 14 candidate rearrangements in 9 of the 10 plasma samples from cancer patients but none in the plasma samples from healthy individuals (Fig. 3). These rearranged sequences were evaluated further by polymerase chain reaction (PCR) amplifications across the rearrangement junctions in tumor and normal DNA from the same nine cancer patients, and all were confirmed to be present in the tumor samples but not in the matched normal DNA. Independent sequencing of the rearranged regions identified the expected rearrangement junctions in all nine cases analyzed. We further evaluated the specificity of the approach by analyzing more than 5.6 billion paired-end Illumina reads of normal

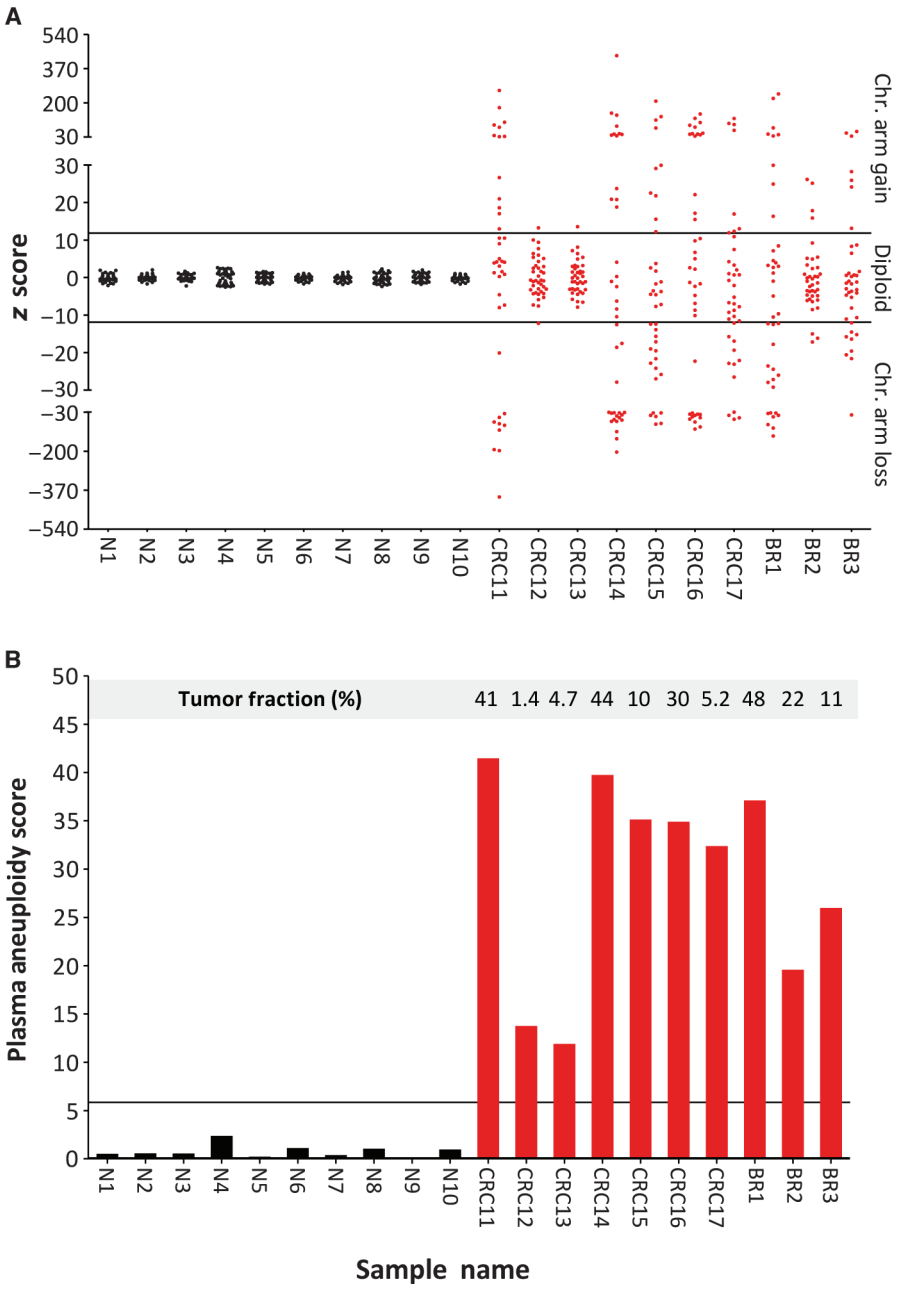
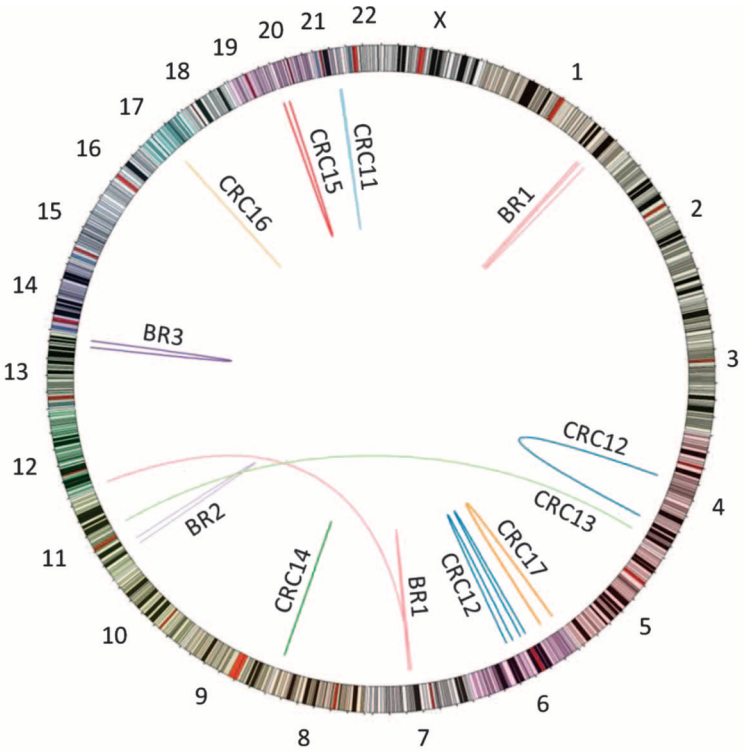



**Fig. 2.** Copy number analysis of plasma samples. (**A**) The *z* scores for each chromosome arm indicate the number of SDs from the mean of the mapped read fraction of the plasma DNA from unaffected individuals (N1 to N10). Positive *z* scores indicate chromosome gains, whereas negative *z* scores indicate chromosome losses. Significant chromosome arm gains and losses were observed only in plasma samples from patients with cancer (CRC11 to CRC17 and BR1 to BR3). (**B**) The PA score was calculated as the number of SDs from the mean of the sum of the −log of the *P* values for the top five chromosome *z* scores of the 10 reference samples (N1 to N10). A PA score of 5.84 (horizontal line) was estimated to indicate aneuploidy in the plasma sample at a specificity greater than 99% (Student's *t* distribution) (see Materials and Methods).




| Sample | Plasma tumor fraction (%) | Rearrangement type | Chr 1 | Breakpoint position 1 | Chr 2 | Breakpoint position 2 | Gene(s) associated with rearrangement* | Breakpoint description* |
|---|---|---|---|---|---|---|---|---|
| CRC11 | 41.3 | Intrachromosomal, inversion | 21 | 32,610,787 | 21 | 34,189,321 | URB1 | Intragenic |
| | | Intrachromosomal, deletion | 21 | 33,757,071 | 21 | 35,045,637 | TMEM50B | Intragenic |
| | | Intrachromosomal, deletion | 21 | 35,145,207 | 21 | 36,324,769 | RUNX1 | Intragenic |
| CRC12 | 1.4 | Intrachromosomal, inversion | 4 | 81,213,132 | 4 | 155,549,875 | ANTXR2, DCHS2 | Intragenic |
| | | Intrachromosomal, deletion | 6 | 58,080,933 | 6 | 67,229,823 | - | Intergenic |
| | | Intrachromosomal, deletion | 6 | 82,048,796 | 6 | 93,604,709 | - | Intergenic |
| CRC13 | 4.7 | Interchromosomal | 11 | 68,718,966 | 4 | 180,111,536 | - | Intergenic |
| CRC14 | 43.8 | Intrachromosomal, duplication | 8 | 141,375,810 | 8 | 141,405,769 | TRAPPC9 | Intragenic |
| CRC15 | 10.3 | Intrachromosomal, inversion | 20 | 1,249,342 | 20 | 11,067,778 | SDCBP2 | Intragenic |
| CRC16 | 29.9 | Intrachromosomal, inversion | 17 | 35,163,714 | 17 | 35,277,642 | ERBB2 | Amplification (10.5-fold) |
| CRC17 | 5.2 | Intrachromosomal, inversion | 6 | 3,938,060 | 6 | 28,043,040 | - | Intergenic |
| BR1 | 47.9 | Intrachromosomal, deletion | 1 | 202,785,323 | 1 | 203,144,644 | MDM4, NFASC | Intragenic |
| | | Intrachromosomal, inversion | 1 | 207,834,294 | 1 | 218,544,458 | CAMK1G | Intragenic |
| | | Interchromosomal | 7 | 86,870,718 | 12 | 3,842,273 | CDK6 | Amplification (6.5-fold) |
| | | Intrachromosomal, deletion | 7 | 87,201,404 | 7 | 90,916,826 | CDK6 | Amplification (6.5-fold) |
| BR2 | 21.5 | Intrachromosomal, deletion | 11 | 24,064,863 | 11 | 34,400,397 | - | Intergenic |
| BR3 | 11.0 | Intrachromosomal, inversion | 13 | 97,936,019 | 13 | 109,030,866 | STK24 | Intragenic |

*The gene disrupted by the rearrangement breakpoint is indicated. If the rearrangement is associated with a focal amplification within 1.5 Mb of the target gene, then the presumed target of the amplification is indicated.

**Fig. 3.** Detection of tumor-specific rearrangements in plasma samples. The Circos plot at the top indicates the rearrangements identified in plasma samples from cancer patients (CRC11 to CRC17 and BR1 to BR3). The type and individual boundaries of the rearrangements are indicated in the lower table. No rearrangements were identified in plasma samples from unaffected individuals (N1 to N10). Rearrangements listed for sample CRC12 were identified in tumor DNA and confirmed in patient plasma, whereas those listed for all other samples were identified directly from patient plasma.

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.