Non-invasive multi-cancer detection using DNA hypomethylation of LINE-1 retrotransposons

Marc Michel, Maryam Heidary, Anissa Mechri, Kevin da Silva, Marine Gorse, Victoria Dixon, Klaus von Grafenstein, Charline Bianchi, Caroline Hego, Aurore Rampanou, et al.

Noninvasive Multicancer Detection Using DNA Hypomethylation of LINE-1 Retrotransposons

Marc Michel[1,2,3,4], Maryam Heidary[4], Anissa Mechri[1,5], Kévin Da Silva[5], Marine Gorse[5], Victoria Dixon[5], Klaus von Grafenstein[5], Charline Bianchi[5], Caroline Hego[4], Aurore Rampanou[4], Constance Lamy[6], Maud Kamal[6], Christophe Le Tourneau[6], Mathieu Séné[7], Ivan Bièche[7], Cécile Reyes[8], David Gentien[8], Marc-Henri Stern[9], Olivier Lantz[10,11], Luc Cabel[12,13], Jean-Yves Pierga[4,12,14], François-Clément Bidard[4,12,15], Chloé-Agathe Azencott[2,3], and Charlotte Proudhon[1,4,5]

ABSTRACT

Purpose: The detection of ctDNA, which allows noninvasive tumor molecular profiling and disease follow-up, promises optimal and individualized management of patients with cancer. However, detecting small fractions of tumor DNA released when the tumor burden is reduced remains a challenge.

Experimental Design: We implemented a new, highly sensitive strategy to detect bp resolution methylation patterns from plasma DNA and assessed the potential of hypomethylation of long interspersed nuclear element-1 retrotransposons as a noninvasive multicancer detection biomarker. The Detection of Long Interspersed Nuclear Element Altered Methylation ON plasma DNA method targets 30 to 40,000 young long interspersed nuclear element-1 retrotransposons scattered throughout the genome, covering about 100,000 CpG sites and is based on a reference-free analysis pipeline.

Results: Resulting machine learning–based classifiers showed powerful correct classification rates discriminating healthy and tumor plasmas from six types of cancers (colorectal, breast, lung, ovarian, and gastric cancers and uveal melanoma, including localized stages) in two independent cohorts (AUC = 88%–100%, $N = 747$). The Detection of Long Interspersed Nuclear Element Altered Methylation ON plasma DNA method can also be used to perform copy number alteration analysis that improves cancer detection.

Conclusions: This should lead to the development of more efficient noninvasive diagnostic tests adapted to all patients with cancer, based on the universality of these factors.

Introduction

Extensive research has shown that tumor genetic alterations can be detected from plasma DNA of patients with cancer (1–3). This paved the way for the use of molecular analyses performed from *liquid biopsies* to genotype tumors noninvasively (4, 5) and demonstrated the potential of ctDNA as a marker of cancer progression (6, 7). It is also a powerful prognostic factor (8) enabling detection of tumor masses not perceptible clinically, after surgery or during treatment. These approaches promise optimal management of patients with cancer and are currently playing an important role in oncology (9, 10). However, several technological obstacles still limit their widespread application. Samples collected at early stages of tumor progression, or during and after treatment, may contain less than one mutant copy per milliliter of plasma (1, 11). This is below the detection limit of most used technologies, even when testing multiple genetic alterations simultaneously. Moreover, most methods are biased toward preselected recurrent mutations, which do not cover all tumors. We observed in our previous studies (12–15) that approximately 25% of patients affected with breast cancer do not display common mutations trackable in plasma DNA, even at advanced stages. Therefore, it is necessary to develop more sensitive and more informative detection tools.

Multiple studies have demonstrated the central role of epigenetic processes in the onset, progression, and treatment of cancer. Epigenetic alterations (i.e., changes in the pattern of chromatin modifications such as DNA methylation and histone modifications) are promising candidates for cancer detection, diagnosis, and prognosis (16, 17). These *extended* markers provide an additional level of information, overlooked by methods that only question genetic alterations (18). Aberrant DNA methylation is a hallmark of neoplastic cells (16), which combine hypermethylation of a wide range of tumor suppressor genes along with a global hypomethylation of the genome (19). DNA methylation is a stable modification, which affects a large number of CpG sites per region and per genome and

[1]Inserm U934, CNRS UMR3215, Institut Curie, PSL Research University, Paris, France. [2]CBIO-Center for Computational Biology, Mines Paris, PSL Research University, Paris, France. [3]INSERM U900, Institut Curie, PSL Research University, Paris, France. [4]Circulating Tumor Biomarkers Laboratory, INSERM CIC BT-1428, Institut Curie, Paris, France. [5]Univ Rennes, Inserm, EHESP, Irset (Institut de Recherche en Santé, Environnement et Travail) - UMR_S 1085, Rennes, France. [6]Department of Drug Development and Innovation (D3i), Institut Curie, Paris, France. [7]Pharmacogenomics Unit, Genetics Department, Institut Curie, Paris, France. [8]Genomics Platform, Translational Research Department, Research Center, Institut Curie, PSL Research University, Paris, France. [9]Inserm U830, Institut Curie, PSL Research University, Paris, France. [10]Inserm U932, Institut Curie, PSL Research University, Paris, France. [11]Laboratory of Clinical Immunology, INSERM CIC BT-1428, Institut Curie, Paris, France. [12]Department of Medical Oncology, Institut Curie, Paris and Saint Cloud, France. [13]CNRS UMR144, Institut Curie, PSL Research University, Paris, France. [14]Université Paris Cité, Paris, France. [15]UVSQ, Université Paris-Saclay, Saint Cloud, France.

M. Michel and M. Heidary contributed equally to this article.

Corresponding Author: Charlotte Proudhon, Institut de Recherche en Santé, Environnement et Travail (IRSET), 9 Avenue du Pr Léon Bernard, Rennes 35000, France. E-mail: charlotte.proudhon@inserm.fr

Clin Cancer Res 2024;XX:XX–XX

doi: 10.1158/1078-0432.CCR-24-2669

Translational Relevance

The Detection of Long Interspersed Nuclear Element Altered Methylation ON plasma DNA assay is a new, highly sensitive strategy to detect bp resolution methylation patterns of long interspersed nuclear element-1 retrotransposons from plasma DNA. It targets 30 to 40,000 young long interspersed nuclear element-1 retrotransposons scattered throughout the genome, covering about 100,000 CpG sites, and is based on a reference-free analysis pipeline. This provided high coverage data using affordable sequencing depth, which is instrumental to achieve high sensitivity and work with minute amounts of cell-free DNA. Resulting machine learning–based classifiers showed powerful discrimination between healthy and tumor plasmas from six types of cancers (colorectal, breast, lung, ovarian, and gastric cancers and uveal melanoma, including localized stages) in two independent cohorts (AUC = 88%–100%, $N = 747$). The Detection of Long Interspersed Nuclear Element Altered Methylation ON plasma DNA data can also be used to perform copy number alteration analysis that improves cancer detection.

will be key to achieve increased detection sensitivity (20). Moreover, the concordance of the methylation status between multiple CpGs of the same region can help detect low frequency anomalies among a heterogeneous population of molecules (21, 22). Finally, combining several genomic regions allows to capture a wide range of tumor alleles and cover the heterogeneous profiles of patients with cancer (23).

Previous studies have shown that cellular DNA methylation patterns are conserved in cell-free DNA (cfDNA) and that detection of cancer-specific profiles at the genome-wide scale is feasible (24–27). Until now, most studies investigating plasma DNA methylation patterns have targeted a limited number of regions at high depth, using PCR-based methods (28–30), or explored genome-wide at low depth with high-throughput sequencing (24–26, 31). Both approaches have limited sensitivity, as focusing on a few regions does not cover cancer-type and patient variability and low depth cannot detect small fractions of ctDNA. More recent studies, relying on the capture of regions of interest coupled with deep sequencing, have investigated the performance of larger numbers of regions at high depth (21, 32–40). These methods enabled sensitive detection and classification of cancer from plasma DNA. However, because they largely focus on cancer hypermethylation and unique sequences, it involves targeting specific regions for each cancer type. As a result, developing a cost-effective universal pan-cancer test remains a challenge.

Remarkably, cancer-related hypomethylation has been reported in almost all classes of repeated sequences (41), from dispersed retrotransposons to clustered satellite repeated DNA, and within multiple forms of cancers (42). In particular, this leads to the reactivation of retrotransposons, resulting in the acquisition of genomic instability, chromosomal rearrangements, and the production of chimeric transcripts between the transposable element and its adjacent locus. Hypomethylation of the internal promoter of long-interspersed element-1 (LINE-1) retrotransposons (L1) has been described as a hallmark of many human cancers (42, 43), which can result in the reactivation of intact L1 elements (44) and the abnormal production of their transcripts and proteins. Transposition of these competent elements induces DNA double-strand breaks and damages the genome. A recent study identified four types of cancer (esophagus, head and neck, lung, and colorectal) with a large amount of damage linked to retrotranspositions involving mostly L1s (45). Another study identified the transposition event responsible for initiating colorectal cancer by mutating the *APC* gene (46). To obtain a global representation of the hypomethylation occurring during carcinogenesis and to increase sensitivity, we chose to target primate-specific copies of L1 retrotransposons (L1PA). These elements have tens of 1,000 of copies per cell and are hypomethylated in multiple cancers (42). Two studies have explored L1 global methylation profiles from plasma (47, 48) of lung and colorectal cancers, using qPCR-based methods, but reported a low detection sensitivity, below 70%. Indeed, repeats being inherently difficult to map, detecting their methylation profiles at the single bp resolution requires sophisticated downstream analysis.

To overcome this, we have developed a method to detect methylation patterns of primate-specific L1 elements (L1PA) from cfDNA, which we named Detection of Long Interspersed Nuclear Element Altered Methylation ON plasma DNA (DIAMOND). We implemented computational tools to accurately align sequencing data without a reference genome and applied prediction models, trained by machine learning algorithms, integrating patterns of methylation, overall and at the single molecule level. The aim of this study was to assess the potential of circulating DNA methylation changes at L1s as a universal tumor biomarker, and to develop new highly sensitive strategies to detect cancer-specific signatures in blood.

Materials and Methods

Cell lines

Cell lines tested in the study are the following: CRC (HCT116 RRID: CVCL_0291); OVC (SKOV RRID: CVCL_0532, Caov3 RRID: CVCL_0201, ES-2 RRID: CVCL_3509); BRC (MDA-MB453 RRID: CVCL_0418, SKBR3 RRID: CVCL_0033, MDA-MB361 RRID: CVCL_0620, HCC202, ZR75.1 RRID: CVCL_0588, HCC70 RRID: CVCL_1270, BT474 RRID: CVCL_0179, MDA-MB231 RRID: CVCL_0062, Cal51 RRID: CVCL_1110, MDA-MB157 RRID: CVCL_0618, BT20 RRID: CVCL_0178, MCF7 RRID: CVCL_0031, HCC1954 RRID: CVCL_1259, HCC1569 RRID: CVCL_1255, HCC38 RRID: CVCL_1267); and UVM (MP38 RRID: CVCL_4D11, MP41 RRID: CVCL_4D12, MP46 RRID: CVCL_4D13, MP65 RRID: CVCL_4D14, MM28 RRID: CVCL_4D15, Mel285 RRID: CVCL_C303, Mel270 RRID: CVCL_C302, 92.1 RRID: CVCL_8607, Mel202 RRID: CVCL_C301, omm2.5 RRID: CVCL_C307, Mel290 RRID: CVCL_C304, mm66 RRID: CVCL_4D17, omm1 RRID: CVCL_6939).

Tissue and plasma samples

Archived tissue samples (ovarian adjacent tumor tissues, ovarian primary and metastatic tumors, breast tumors, and uveal melanoma tissues) were retrieved from the Pathology Department of Institut Curie. Healthy white blood cells and healthy plasma were collected from blood of healthy donors through the French blood establishment (agreement #16/EFS/031) under French and European ethical practices. Blood samples from patients treated at the Institut Curie were collected, after written informed consent, as part of the following studies: resectable metastatic colorectal cancers from the Prodige14 trial (approved by a French Personal Protection

Committee – "CPP—Comité de Protection des Personnes Sud Méditerranée IV" and registered in ClinicalTrials.gov under NCT01442935); non–small cell lung cancer and metastatic HR^+ $HER2^-$ breast cancer from the ALCINA study (approved by a French Personal Protection Committee and registered in ClinicalTrials.gov under NCT02866149); treatment-naïve patients with ovarian cancer or triple-negative breast cancer (TNBC) eligible for surgery or neoadjuvant chemotherapy from the SCANDARE study (approved by the French National Agency for the Safety of Medicines and Health Products "ANSM—Agence National de Sécurité du Médicament," a French Personal Protection Committee and registered in ClinicalTrials.gov under NCT03017573); multiple types of metastatic cancers from the SHIVA02 study (approved by the French National Agency for the Safety of Medicines and Health Products "ANSM—Agence National de Sécurité du Médicament," a French Personal Protection Committee and registered in ClinicalTrials.gov under NCT03084757); and nonmetastatic operable gastric cancers and advanced uveal melanoma from CTC-CEC-ADN study (approved by a French Personal Protection Committee and registered in ClinicalTrials.gov under NCT02220556). Additional archived samples were also retrieved from the biobank of the Institut Curie, patients having provided informed consent for research use. All samples were obtained in accordance with the ethical guidelines, with the principles of Good Clinical Practice and the Declaration of Helsinki. This study was approved by the Internal Review Board and Clinical Research Committee of the Institut Curie. Blood samples were collected at the time of inclusion, before the start of the treatment, in EDTA tubes. Plasma was isolated within 4 hours, to ensure a good quality of cfDNA, by centrifugation at 820 *g* for 10 minutes, followed by a second centrifugation of the supernatant at 16,000 *g* for 10 minutes and stored at −80°C until use.

Preparation of DNA from cell lines and tissues and cfDNA

Isolation of DNA from cell lines and healthy white blood cells (buffy coats) was performed using the QIAamp DNA Mini Kit or QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's instructions. DNA from cryopreserved and formalin-fixed paraffin embedded tumor tissues was extracted using a classical phenol chloroform protocol and the NucleoSpin FFPE DNA kit (Macherey Nagel), respectively.

cfDNA was extracted from 2 mL of plasma using the automated QIAsymphony Circulating DNA Kit (Qiagen), the Maxwell RSC ccfDNA LV Plasma Kit (Promega), or manual QIAamp Circulating Nucleic Acid Kit (Qiagen), according to the manufacturer's instructions, and eluted in 60, 75, or 36 μL, respectively. We verified that the extraction method did not impact our results (Supplementary Fig. S1A and S1B). Isolated DNA was quantified by Qubit 2.0 Fluorometer using dsDNA HS Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions and stored at −20°C until use.

Bisulfite conversion

We used sodium bisulfite–based chemical conversion to achieve bp resolution analysis, which is crucial to address methylation levels at single CpG dinucleotides and the co-methylation of multiple CpG sites to determine methylation *haplotypes* (methylation state of successive CpG sites). Bisulfite treatment of the isolated genomic DNA (up to 200 ng) from the cancer tissues, cancer cell lines, and buffy coats was performed using an EZ DNA Methylation-Gold Kit (Zymo Research), following the manufacturer's instructions. Bisulfite treatment of cfDNA (isolated from 2 mL of plasma) was performed using the Zymo EZ DNA Methylation-Lightning Kit (Zymo Research), according to the manufacturer's instructions. Bisulfite-treated DNA was stored at −80°C and further used to build a sequencing library. We have compared the methylation profiles obtained with bisulfite conversion or enzymatic conversion [NEBNext Enzymatic Methyl-seq Conversion Module, (49)] followed by amplification with the DIAMOND targets and deep sequencing. We observed similar methylation profiles (Supplementary Fig. S2). We have tested various starting quantities of DNA (100, 10, and 5 ng) of white blood cells extracted from healthy donors (buffy coats). We have also compared cfDNA (10 ng) and DNA from MCF7 breast cancer cell lines (100 ng), known to display hypomethylation at L1PA elements. We observed that the methylation profiles along the 30 CpG targets are similar when starting with 100 ng (Supplementary Fig. S2A) but, in our hands, the bisulfite conversion seems more robust with decreasing stating DNA quantities (Supplementary Fig. S2B and S2C). We thus observe some differences with the cfDNA samples at 10 ng (Supplementary Fig. S2D). We could still detect very similar hypomethylation profiles in MCF7 breast cancer cell lines (Supplementary Fig. S2E). This shows that EM-seq could be used to profile L1PA methylation changes but the same conversion should be used throughout the study.

Primer design

Eight primer pairs were designed using the LINE-1 human-specific (L1HS) consensus sequence from Repbase (RRID: SCR_021169; **Fig. 1A**). Although 5′ untranslated region (promoter region) is CpG-rich and common target for methylation quantitation, L1PA copies are frequently 5′-truncated. Therefore, primers were also designed for ORFI and ORFII to target more L1PA elements and improve the sensitivity of our assay. All primers were designed for plus strand of bisulfite-converted DNA, using the MethPrimer (RRID: SCR_010269) or PyroMark software (RRID: SCR_018617). Targeted regions contained 2 to 7 CpG targets and ranged from 101 bp to 150 bp, to better capture cfDNA fragments, which have a mean size of 167 bp (50), (Supplementary Table S1). Primers were methylation-independent, encompassing 0 to 2 CpGs (none toward the 5′ end), and were degenerated to target both the methylated and unmethylated states. They contained Fluidigm universal common sequence (CS) tags at their 5′ ends. We incorporated a 16 N (random nucleotides) as unique molecular identifiers (UMI) between the target-specific sequence and the CS2 in the reverse primers for signal deconvolution to detect true low frequency alterations and for reducing errors. As LINE-1 hold 1,000 of copies per genome, a high number of distinct UMIs are essential for unique barcoding of each target molecule. The 16 N stretch between the target-specific sequence and the CS1 in forward primers was used to increase the diversity of sequencing libraries and improve sequencing quality. All primers were obtained from Eurogentec (RP-cartridge purification method).

Preparation of targeted bisulfite sequencing libraries

To limit batch effects, sequencing and library preparation batches contained both cancer samples and healthy donors. We also specifically processed the validation cohort C2 with equilibrated number of healthy and cancer samples distributed in five experimental batches. Sequencing libraries were prepared using three PCR steps (Supplementary Fig. S3A): (i) target-specific linear amplification for UMI assignment, (ii) target-specific exponential amplification, and (iii) barcoding PCR for sample identification. Each library was prepared

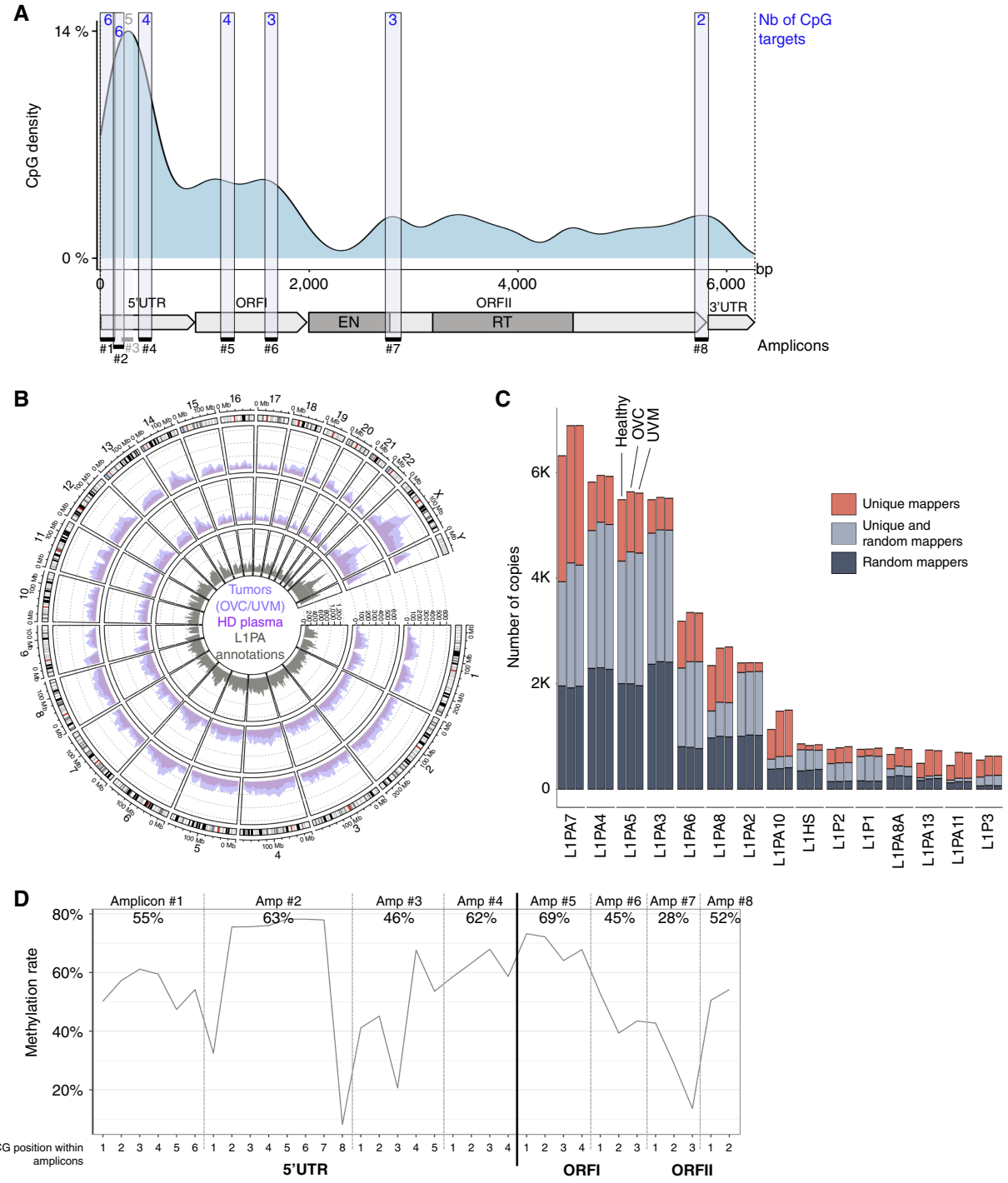

Figure 1.

Targeting primate-specific LINE-1 elements reveals genome-wide plasma DNA methylation patterns. **A,** CpG density along the structure of an L1HS element, which contains 95 CpG. The DIAMOND assay targets 30 CpG. Each target amplicon is highlighted by a black bar below the structure. The number of CpG sites detected per amplicon is displayed in blue. **B,** L1PA copy number hit by uniquely and/or randomly mapped reads, obtained from a healthy plasma vs. ovarian (OVC, top track) or uveal melanoma (UVM, middle track) tumor tissue samples "deep sequenced" (54M, 44M, or 46M reads, respectively) over the distribution of L1PA elements annotated in the genome (RepeatMasker on hg38, gray bottom track). **C,** Histogram summarizing the most represented subfamilies of L1 targeted by the DIAMOND assay in the three "deep-sequenced" samples, in descending order (sum of copies across the three samples). The colors highlight the relative contribution of L1PA copies hit by reads uniquely mapped, randomly mapped, or both. **D,** Methylation pattern observed across the eight regions targeted along the L1 element in the healthy plasma sample "deep-sequenced." Metaplot showing the average methylation levels at each CpG position. Amplicon limits are delineated with gray dotted lines. The dark line marks the end of the 5′UTR. Average levels per amplicon are indicated. UTR, untranslated region.

in two individual reactions (due to the overlap of amplicon 2 with other primers), including (i) multiplex PCR amplification of seven probes (amplicons 1, 3, 4, 5, 6, 7, and 8), and (ii) single PCR amplification of amplicon 2.

UMI assignment for multiplex reaction was performed using Platinum Multiplex PCR Kit Master Mix (Thermo Fisher Scientific, Life Technologies SAS) in a 25 μL reaction containing 1× Platinum Multiplex PCR Master Mix, 0.01 to 0.06 μmol/L mix of reverse primers and up to 5 ng bisulfite-converted DNA at the following thermocycling conditions: 95°C for 5 minutes followed by one cycle at 95°C for 30 seconds, 58°C for 90 seconds, and 72°C for 40 seconds. UMI assignment for single reaction was performed using Hot Star Taq Plus DNA Polymerase (Qiagen) in a 25 μL reaction containing 1× Taq PCR Buffer, 0.65 U Hot Star Taq (5 U/μL), 0.2 μmol/L dNTPs, 1.5 mmol/L $MgCl_2$, 0.1 μmol/L amplicon 2 reverse primer, and up to 4 ng of bisulfite-converted DNA at the following thermocycling conditions: 95°C for 10 minutes followed by one cycle at 94°C for 60 seconds, 58°C for 30 seconds, and 72°C for 40 seconds. To ensure complete removal of the reverse primers and dNTPs, each 25 μL reaction was treated with 50 U of Exonuclease I and 10 U of FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific) at 37°C for 1 hour and heat-inactivated at 80°C for 15 minutes.

Target-specific exponential amplification for multiple reaction was performed using Platinum Multiplex PCR Kit Master Mix in a 50 μL reaction containing 1× Platinum Multiplex PCR Master Mix, 0.01 to 0.06 μmol/L mix of forward primers, 0.2 μmol/L CS2 reverse primer, and 20 μL of purified PCR product at the following thermocycling conditions: 95°C for 5 minutes followed by 28 cycles at 95°C for 30 seconds, 58°C for 90 seconds, and 72°C for 30 seconds followed by a 10-minute incubation at 72°C. Target-specific exponential amplification for single reaction was performed using Hot Star Taq Plus DNA Polymerase in a 25 μL reaction containing 1× Taq PCR Buffer, 0.65 U Hot Star Taq (5 U/μL), 0.2 μmol/L dNTPs, 1.5 mmol/L $MgCl_2$, 0.2 μmol/L amplicon 2 forward primer, 0.2 μmol/L CS2 reverse primer, and 8 μL of purified PCR product at the following thermocycling conditions: 95°C for 10 minutes, 25 cycles at 94°C for 60 seconds, 58°C for 30 seconds, 72°C for 30 seconds, and 10 minutes at 72°C.

PCR products of multiplex and single reaction were pooled together after quantification by qPCR and purified using Agencourt AMPure XP (Beckman Coulter) at 1.2× ratio according to the manufacturer's protocol. Purified DNA was eluted in 30 μL of water. Barcoding PCR was performed using universal Fluidigm primers. Purified pooled PCR product (25 μL), 1× Phusion HF Buffer, 1 U Phusion Hot Start II DNA Polymerase (Thermo Fisher Scientific), 0.2 μmol/L Fluidigm primer, and 0.2 mmol/L dNTPs were mixed in the final volume of 50 μL and amplified with the following conditions: 98°C for 2 minutes, followed by 20 to 25 cycles of 98°C for 10 seconds, 62°C for 30 seconds, and 72°C for 30 seconds followed by a 5-minute incubation at 72°C. The amplified product was purified by two consecutive AMPure XP steps using (i) a low concentration of AMPure XP beads (0.6× to 0.7× ratio) in which the beads containing the larger fragments are discarded and supernatant collected (reverse purification) and (ii) higher beads concentration (1.1× to 1.2× ratio) in which the beads containing fragments of interest were collected and purified according to the manufacturer's protocol. Size-selected libraries were eluted in 15 μL of low-EDTA TE buffer. The libraries were quantified with Qubit HS DNA kit (Thermo Fisher Scientific), qualified with nano-electrophoresis (TapeStation, Agilent RRID: SCR_014994), and pooled equimolarly for sequencing. Sequencing was performed on Illumina HiSeq rapid run mode or NovaSeq (PE 30 bp, 170 bp).

Preprocessing of the reads

For each sample, FASTQ files containing raw sequences, composed by the following parts: CS1, forward UMI, forward primer, insert, reverse primer, reverse UMI, and CS2 (Supplementary Fig. S3A) were first filtered for reads quality (average >Q20 per read) and then demultiplexed (i.e., cut using *atropos* v1.1.31 RRID: SCR_023962) using forward and reverse primer sequences. FASTA files were created per primer-set, containing inserts and reverse UMIs for deduplication, as they are unique for each input DNA molecule. Inserts and reverse UMI were then filtered on expected sizes (with a tolerance of ± 5 bases for the inserts). Filtered inserts and UMIs sequences were concatenated and deduplicated using *vsearch* v2.15.2 (RRID: SCR_024494). Reverse UMIs were then trimmed and resulting inserts from all samples were aggregated into a single FASTA file per primer-set.

Clustering, extraction of representative sequences, and global alignment

Using *vsearch* (with the following parameters: –cluster_fast <inputFasta> –notrunclabels –fasta_width 0 –iddef 4 –id 0 –qmask none –clusterout_sort –consout <referenceFasta>), a clustering based on sequence identity was applied to each FASTA file, or a subset of 20 million reads randomly chosen if a given file comprised more. The 10 largest clusters' representative sequences were isolated in separate files. Using MAFFT v7.508 (RRID: SCR_011811; with the following parameters: –globalpair –maxiterate 1000), the 10 representative sequences were aligned pairwise resulting in a reference database for each primer-set. Lastly, using *mothur v*1.48.0 (RRID: SCR_011947, with the following parameters: #align.seqs(candidate=<inputFasta>, template=<referenceFasta>, align=needleman, match=1, mismatch=−1, gapopen=−1, gapextend=0)) on each primer-set FASTA file, all sequences from all samples were aligned to the corresponding reference.

CG calling, methylation levels, and haplotype extraction

To call CpG dinucleotides of interest, a sliding window of 2 bp was used on all aligned sequences to determine the distribution of dinucleotides along each amplicon target. A first threshold of ≥20% of CG/TG dinucleotides was used to select potential CpG sites. A second threshold was applied to eliminate dinucleotide with ≥95% TG and select position with at least 5% methylation rate. From the aligned sequences, the patterns of methylation were extracted and compiled into either average levels of methylation at each previously identified CpG sites, or proportions of methylation haplotypes for each sample.

Machine learning–based classification models

The resulting data (represented as average levels of methylation per CpG site or proportions of methylation haplotypes or both) were used to do supervised learning of statistical models using the random forest classifier algorithm (51) from Python package scikit-learn (RRID: SCR_002577), with the following hyperparameters: *n*_estimators=300, criterion="gini," max_depth=None, min_samples_split=2, min_samples_leaf=1, minutes_weight_fraction_leaf=0.0, max_features="sqrt," max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None.

The rational for choosing random forest over other learning methods was driven by three main factors: (i) it is less prone to overfitting (51); (ii) it shows excellent performance even when the quantitative relationship between features and observations is biased in favor of the former, such as when using methylation haplotype data representation (52); and (iii) random forests also inherently return measures of variable importance (51), such as mean decrease in impurity, which greatly facilitate the interpretability of model decisions. The features used to train the models were the average levels of methylation per CG site ($n = 30$) and the proportions of methylation haplotypes (i.e., the combinatorial of all the possible methylation status of CG sites within a given amplicon, $n = 372$) or both. No additional transformation nor feature selection was performed on the data.

Expert and all cancer models

Model classifications were run 5,000 times in order to estimate variance and confidence intervals. For the discovery step, in each run, as many samples from each class were randomly drawn to construct a balanced subset of the data (53). The samples from these draws were stratified by class and split into 60% for training, 40% for evaluation. For the validation step, we trained the model on the entire cohort 1 and evaluated it on cohort 2. The true and false positive rates for all possible classification threshold were evaluated at each run, with interpolation to generate an average ROC curve with 95% confidence interval (CI) for the 5,000 runs. Ninety-five percent CI has been calculated with the following formula: $M \pm z \times s/\sqrt{n}$ with M the average of the variable, z the confidence level (z~1.96 for 95% CI), s the SD, and n the number of samples in the variable.

Blind models

We trained a random forest on haplotypes features, removing one cancer type or subgroup from the training set. The specific cancer type or subgroup is then assessed in the test set. We pooled together the discovery and validation cohorts, training on 2/3 of all the samples–excluding the cancer type or subgroup to test for–and testing on the remaining one-third of the samples. The only exception was metastatic gastric cancer: as they are made up of only three samples, they were systematically moved to the test set, consequently blinding the model toward GAC M+ samples. We also trained a stacked version with this setup (see below).

Stacked machine learning model

We developed a model referred to in the article as "stack." This model uses one random forest model for each combination of cancer type and metastatic status, known as the "expert submodel." Each expert submodel was trained on one-third of the healthy plasma samples and one-third of the samples matching the cancer subgroup of interest (cancer type and dissemination status). These expert submodels were then combined into a random forest stack model, which uses both the haplotype features and the probabilities output by each expert submodel. The final random forest stack model was trained on an additional one-third of the healthy and cancer plasma samples and tested on the remaining samples (which represent one-third of the healthy samples and one third of each subgroup).

Mutation screening for ovarian cancer samples

Ovarian tumor genotyping was performed using the TIGER panel previously developed by Institut Curie (54), which targets 78 genes or using "custom next-generation sequencing" with amplicons targeting *Tp53* and *TSC2*, which are the two most frequently mutated genes in this group of patients, with TruSeq library constructions for low input material (dual strand technology). After mutation identification, ovarian cancer plasmas were genotyped using custom next-generation sequencing or Droplet Digital PCR as previously done (14). Sequencing was performed on a MiSeq V3-150 (25M) with paired-end 75bd protocol.

Whole-genome bisulfite sequencing analysis

To see if we could retrieve cancer-associated L1PA hypomethylation in other plasma studies, we analyzed data sets from two recent studies profiling cfDNA methylation with whole-genome bisulfite sequencing (WGBS) in healthy individuals and patients with cancer [Liu and colleagues (55); Gao and colleagues (56)]. Liu and colleagues (55) analyzed methylation at 75,617 CpGs in the whole genome [estimated with Bis-SNP (RRID: SCR_005439) after mapping with Bismark (RRID: SCR_005604)] of 17 healthy donors and 31 patients with cancer. We extracted the methylation levels at CpG residing within L1PA families hit by DIAMOND. To do so, we identified CpG dinucleotides covered by L1HS–L1PA10 elements, and their position within L1 based on the L1HS consensus sequence, and we selected the DIAMOND copies using BEDTools (RRID: SCR_006646). Gao and colleagues (56), provided FASTQ files data for 123 patients with breast cancer and 40 healthy patients. In order to avoid breast cancer subtype effect and age effect, we subsampled the data to generate an age-matched cohort with 16 healthy and 15 HR^+ $HER2^-$ M+ patients. We first merged the paired-end reads using Fastp (RRID: SCR_016962) and mapped them on the hg38 reference genome using Bismark. At this step, we followed two different approaches to obtain the methylation level at the 30 L1PA CpGs that we studied with DIAMOND. (i) Using *Bismark extractor*, we retrieved the methylation levels of all CpGs covered and intersected it with the dinucleotides covered by DIAMOND using BEDTools. Results produced with this approach are subsequently called "Gao and colleagues (56) Bismark." (ii) Using *mothur* (RRID: SCR_011947), we aligned the reads that mapped on L1HS-L1PA10 elements on an aggregate of native and converted (CG converted to TG) reference sequences of L1HS. All reads that mapped with a score lower than 51% of similarity (score computed by *mothur*) were rejected (this threshold was established from the percentile at 99% of similarity score distribution from 1,000 random sequences 167 bp long). Results generated by this approach are subsequently called "Gao and colleagues (56) *mothur*." Finally, for "Gao and colleagues (56) Bismark" and "Gao and colleagues (56) mothur," we refined the results to only take in account the copy of L1HS–L1PA10 targeted by DIAMOND. Further refinements were done to only consider L1HS–L1PA3 copies with the second approach, referred to as "L1HS–L1PA3."

Survival analysis

Survival analysis has been performed with *survival* (RRID: SCR_021137) and *survminer* (RRID: SCR_021094) R (RRID: SCR_001905) packages - r.

Copy number alteration analysis

CytoScan HD microarrays: 250 ng of gDNA from 15 breast cell lines (1 normal-like: HTERT-HME1 and 14 cancer cell lines: MDA-MB231, MDA-MB453, HCC1569, BT20, HCC1954, HCC38, MDA-MB361, ZR 75.1, MDA-MB157, MCF7, SKBR3, HCC202, HCC70, and BT474) were characterized using Affimetrix/Thermo CytoScan

HD microarrays at the Genomics facility of Institut Curie to profile aneuploidy. To compare with the *z*-score by chromosome arm, we calculated the mean of weighted $\log_2$ combining probes by chromosome arms.

DIAMOND copy number alteration (CNA): (i) *Z*-score calculation: prepossessed reads were uniquely mapped on hg38 genome using Bismark (version 0.23.1). As in Belic and colleagues (57), only the reads with an alignment score >15 were kept. Resulting reads from all amplicons (excluding #2 and #3) were merged, and the normalized number of reads per chromosome arm (excluding sexual chromosomes X and Y) per sample was calculated with R. Next, the amplification/deletion score was computed using the following formula:

$$z-score_{i,n} = \frac{ReadsNorm_{i,n} - Mean\left(ReadsNorm_{i,controls}\right)}{Sd\left(ReadsNorm_{i,controls}\right)}$$

with i = a given chromosome arm, n = a given sample, and *controls* = a set of reference samples (white blood cells from 10 healthy reference samples for the cell lines; 63 healthy plasmas from C1 as a reference for cancer and healthy plasma samples). Genome-wide *z*-scores were computed by summing the squared *z*-scores of all chromosome arms. (ii) *Z*-score threshold identification: to identify altered versus normal *z*-scores, we performed 5-fold cross validation of simple cutoff classification model on the discovery cohort ($N_{\text{Healthy}} = 60$, $N_{\text{Cancer}} = 350$) using the genome-wide *z*-score and calculated the threshold that maximize the sensitivity at 100% specificity.

Two-step classification for sample labeling

First, we selected the threshold for the probability of the cancer prediction ($\text{Proba}_{\text{Cancer}}$) on the discovery cohort maximizing the sensitivity for a 99% specificity, per "cancer-type" model. We applied this threshold on the $\text{Proba}_{\text{Cancer}}$ computed with the validation models and reclassified samples which presented a *z*-score > 121, as cancer ($\text{Proba}_{\text{Cancer}} \leq$ Threshold C1 AND GZ-score ≤ 121: prediction = Healthy; $\text{Proba}_{\text{Cancer}} >$ Threshold C1 OR GZ-score > 121: prediction = Cancer).

Data availability

Data have been deposited as methylation matrices (CG % or haplotypes %) on the Zenodo database (RRID: SCR_004129) with the following accession code: https://zenodo.org/uploads/12206227 and as compressed FASTQ files at the European Genome-phenome Archive at https://ega-archive.org/ under the accession code EGAD50000000646. WGBS sequencing data were downloaded from publicly available database at https://zenodo.org/records/7779198 and from the National Center for Biotechnology Information (https://www.ncbi. nlm.nih.gov) under the accession number PRJNA494975. The code used to analyze the data is available on github: https://github.com/ProudhonLab. Source data used to generate figures are available upon request to the corresponding author.

Results

Targeting primate-specific LINE-1 elements reveals genome-wide plasma DNA methylation patterns

We developed a PCR-based targeted bisulfite method coupled to deep sequencing to detect methylation patterns of L1PA elements. We used sodium bisulfite chemical conversion to achieve bp resolution analysis and designed a multiplexed PCR based on eight amplicons covering L1PAs (**Fig. 1A**; Supplementary Fig. S3A; Supplementary Table S1). We detected 1,000 of L1PA elements scattered throughout the genome as shown by the genomic hits obtained from healthy plasma, an ovarian tumor, and a uveal melanoma tumor sequenced at high depth (**Fig. 1B**; Supplementary Table S2). We observed similar profiles for the three samples, as well as for healthy and cancer plasmas with standard coverage (Supplementary Fig. S1B–S1E). This demonstrated the robustness of the approach. Overall, the estimated number of L1PA targets is about 30 to 40,000 elements per genome including half of the human-specific copies (L1HS) and many copies of the other L1PA subfamilies (**Fig. 1C**; Supplementary Table S2). This represents an estimate of 87 to 120,000 CpG sites.

Following deep sequencing, reads are traditionally mapped back to the genome. However, the majority of sequencing reads from repetitive sequences are assigned randomly during mapping steps and are subsequently lost for classical differentially methylated region calling (58). We, thus, developed a new computational pipeline to accurately align repetitive sequencing data without using a reference genome (Supplementary Fig. S3F). To perform this, we clustered all good quality reads based on their similarity, extracted representative sequences from the largest clusters, and used them for multiple sequence alignment. We then aligned all the reads back onto this custom database. Using such reference-free method, we preserved the majority of our data and could extract the informative CpG sites agnostically. We selected sites with a CG/TG content ≥20% including at least 5% of CG to ensure that the position of interest carries some DNA methylation marks. This selection was done on healthy samples to avoid biases related to cancer hypomethylation. We retrieved 35 CpG positions covered by our panel including two additional CpGs with respect to the L1HS consensus annotations, located within amplicon 2 (Supplementary Fig. S4A and S4B).

As expected, the 5′ end of the L1 copies targeted is heavily methylated (42, 59), particularly within the second amplicon. We also observed quite high levels in both the fifth amplicon (69% in average; **Fig. 1D**), which covers part of the ORFI, and the last two CpGs of amplicon 8, which is located immediately upstream of the 3′ untranslated region . Amplicon 3, which has the lowest methylation levels within the 5′ end, displayed sequencing data with atypical distributions and showed less robust performances (not shown). Hence, we further eliminated it from the rest of the study, resulting in a total of 30 CpG positions analyzed. Overall, this reference-free method retrieved methylated sites contained by the youngest LINE-1 elements present in the human genome allowing us to study their DNA methylation levels and motifs from minute amount of DNA such as plasma cfDNA.

L1PA hypomethylation is detectable from plasma DNA in multiple forms of cancer

We first tested the DIAMOND approach on methylation controls, cancer cell lines, and tissue samples. The overall methylation levels demonstrated an extensive L1PA hypomethylation specifically in cancer samples, including colorectal (CRC), ovarian (OVC), breast (BRC), and uveal melanoma (UVM) cancer cell lines as well as ovarian cancer, breast cancer, and uveal melanoma tumors compared with healthy white blood cells and healthy tissues collected adjacent to ovarian tumors (**Fig. 2A**; Supplementary Table S3). Next, we tested a cohort of 473 plasma samples including 123 healthy plasma controls and plasma samples from patients with six different types of cancer, covering metastatic (M+) and localized (M0) stages (Supplementary Table S4). This includes colorectal and

ovarian cancers in which a substantial rate of L1 hypomethylation has previously been reported (60, 61). We detected a statistically significant L1PA hypomethylation in cfDNA of metastatic colorectal cancer (CRC M+), breast cancer (BRC M+), and uveal melanoma (UVM M+) samples as well as in locally advanced ovarian cancers (OVC M0, stages III) and localized gastric cancers (GAC M0; **Fig. 2B**; Supplementary Table S3). The global methylation was not significantly different in metastatic non–small cell lung cancers (LC M+) nor in localized stages of breast cancer (BRC M0). However, focusing strictly on global methylation levels provides only part of the information.

We further computed the levels of methylation at each CpG target ($n = 30$) for these plasma samples and observed specific patterns of methylation along the L1 structure, which are robustly conserved among the 123 healthy donors (**Fig. 2C**). When considering all cancer samples together, we observed a clear difference with the methylation of healthy samples. We detected a steady hypomethylation through all CpG targets except for the two sites within amplicon 8 (**Fig. 2D**). This is also true for metastatic colorectal cancers (CRC M+), breast cancers (BRC M+), and uveal melanoma (UVM M+). Clear hypomethylation is also observable for localized gastric (GAC M0) and ovarian (OVC M0) cancers, in particular at amplicon #1, #4 and #6, while the differences are less striking for localized breast cancers (BRC M0) and metastatic non–small cell lung cancers (LC M+). The distinction between most cancers and healthy samples were dependent on multiple CpG positions belonging to different amplicons along L1s, as shown by principal component analysis (Supplementary Fig. S4C). The least discriminating positions were located within amplicon 8, which is consistent with the metaplots shown in **Fig. 2D**.

Next, we analyzed the motifs of methylation at the molecule level, which provide a more detailed signal. These *haplotypes* correspond to true patterns of methylation of adjacent CpGs, detected for each amplified DNA molecule. This was achieved by the incorporation of UMIs into the library (Supplementary Fig. S3A). Based on the combination of the 30 CpG targets divided into their seven amplicons, we extracted a total of 372 unique features (Supplementary Fig. S4D). We observed highly robust representation profiles of haplotypes among the 123 healthy samples (**Fig. 2E**). For most amplicons, the fully methylated molecules were the most represented, as expected for healthy controls. However, we observed a high proportion of totally unmethylated haplotypes in amplicon #6 and #7. This can be explained by the fact that older L1 copies are often truncated in 5′ and less regulated by DNA methylation, leading to the capture of molecules with lower DNA methylation in 3′. Nevertheless, several intermediate patterns were also among the most important features and were found to be differentially represented in healthy and cancer samples (Supplementary Table S5). Fully methylated haplotypes were significantly under-represented in most cancer subgroups and in most amplicons (**Fig. 2F**). On the contrary, fully unmethylated haplotypes were over-represented in most cancer subgroups and in most amplicons. This is also well illustrated by the principal component analysis shown in Supplementary Fig. S4E, underlining the contribution of highly methylated haplotypes toward the healthy group versus the lowly methylated haplotypes separating cancer samples (middle). This separation involves haplotypes from all amplicons (right).

Next, we compared the methylation profiles of tumor and plasma paired samples (OVC = 10, Supplementary Fig. S5A–S5F; BRC = 16, Supplementary Fig. S5G–S5K) by calculating the correlation between their methylation differences relative to the mean methylation of healthy donor plasmas. We observed a better correlation with methylation haplotype portions than with single CG methylation features (Supplementary Fig. S5B, S5C, S5H, and S5I; Supplementary Table S6). These results demonstrate that L1 hypomethylation can robustly be observed from cancer plasma DNA at the level of single CpG sites and more importantly at the level of methylation haplotypes.

L1PA hypomethylation–based classifiers recognize samples from multiple forms of cancer

We then trained classification models using random forests, with the 30 features corresponding to the levels of methylation at each CpG target or the 372 features corresponding to the proportions of haplotypes, or both, and assessed their performances to automatically separate healthy from tumor plasmas. By testing all cancer samples without cancer-type specification, the methylation of L1PA elements showed an extremely good ability to discriminate between healthy and tumor plasmas, with an overall AUC of 94% to 95% for the three types of features (**Fig. 3A** and **C**; Supplementary Fig. S6A and S6B). Next, we trained distinct models to estimate the performances for each cancer type and/or dissemination stage (M0 vs. M+). These models were extremely performant in metastatic colorectal and breast cancers but also stage III ovarian cancers and localized gastric cancers, with nearly perfect classifications and AUCs between 98% and 100% (**Fig. 3B** and **C**; Supplementary Fig. S6A and S6B). Additionally, we observed excellent performances for metastatic lung cancers and uveal melanoma and more importantly for localized stages of breast cancer ($AUC_{BRC\_M0} = 92\%–95\%$). These models provide very good sensitivities at 99% specificity (**Fig. 3D**), in particular for CRC M+, BRC M+, OVC M0, GAC M0, and BRC M0. The latter is one of the most difficult cancers to detect noninvasively, as reported in previous liquid biopsy multicancer tests (11, 39, 40).

Overall, we observed similar results using single CpG methylation levels, haplotype proportions, or both features. This can be explained by the high correlation observed between the two types of features (Supplementary Fig. S6C). We compared these performances for each subgroup with the classification rates extracted from the model "all," testing all cancer samples together. We observed similar AUCs for most cancers, but overall, they were better classified with their "expert" cancer-specific models (Supplementary Fig. S6D and S6E). This shows that certain specificities of cancer type may confuse the current "all" model and affect the sensitivities at 99% specificities.

Subsequently, we evaluated the importance of the features used by our classifiers (**Fig. 3E** and **F**). CpG positions displayed different patterns in the various cancer subgroups that can be informative for distinct cancer types or stages (**Fig. 3E**). Nonetheless, we identified features which are common to many types of cancer such as most CpGs of amplicon 1 and the first CpG of amplicon 6. Other features seemed to be characteristic of specific subgroups, such as CG7-14 which are the most important features for sorting localized stages of breast cancer (BRC M0) or CG15-18, in particular CG17, which are part of the top features for metastatic breast cancers (BRC M+). Besides their dissemination status, the different important features detected in the BRC M0 and M+ subgroups may result from the fact that they were composed of different breast cancer subtypes, with 100% of hormone dependent (HR^+ $HER2^-$) cancers in the BRC M+ subgroup and a mixture of subtypes, including also TNBC and $HER2^+$ cancers, in the BRC M0 subgroup (Supplementary Table S7). Indeed, we observed that slightly different hypomethylation

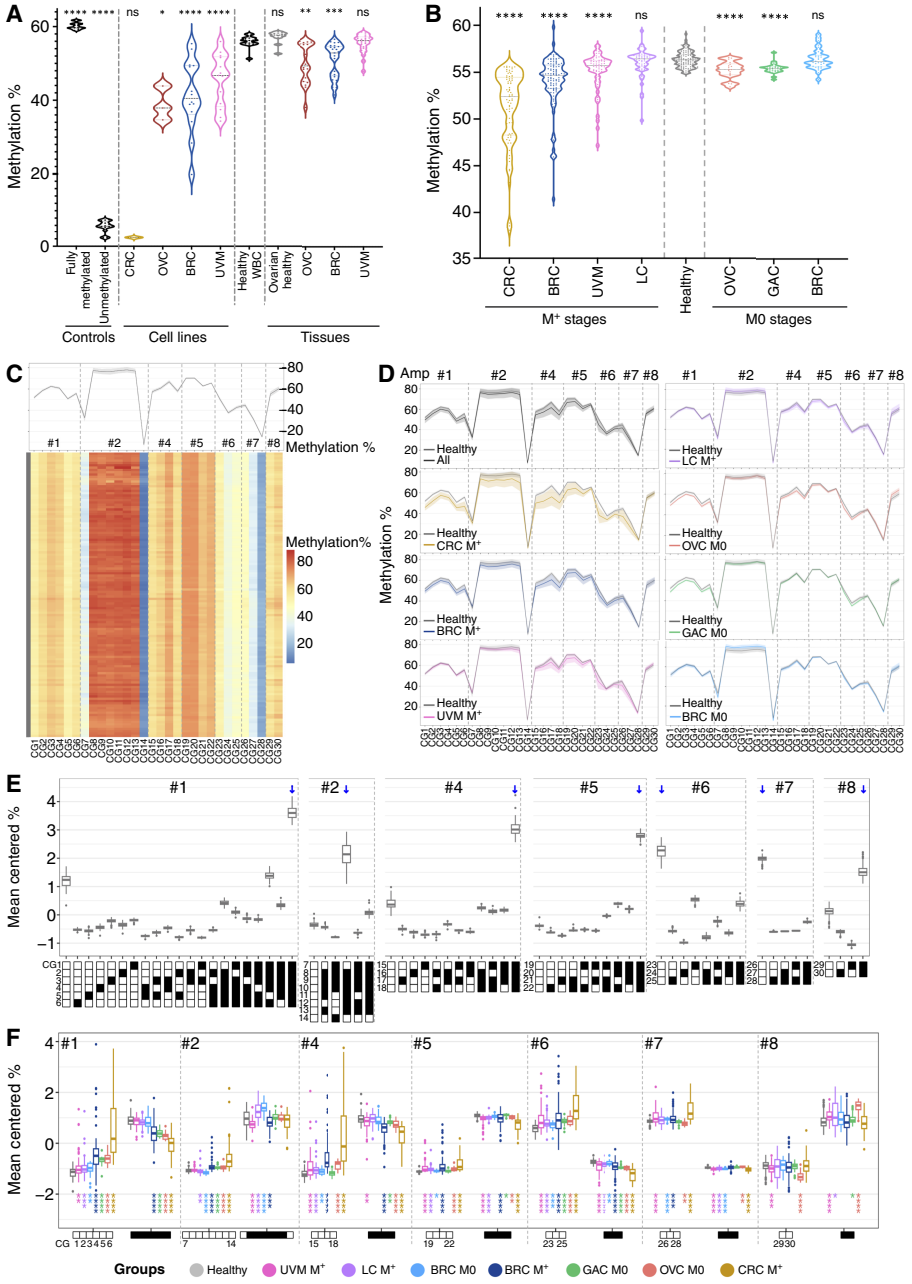

Figure 2.

L1PA hypomethylation is detectable from plasma DNA in multiple forms of cancer. **A,** Global DNA methylation of fully methylated (healthy WBC DNA treated with SssI) and unmethylated (whole-genome amplified healthy WBC DNA) controls, cancer cell lines, or tissues. Ovarian healthy tissues were collected next to ovarian tumors. The global methylation levels for each sample correspond to the percentage of CG dinucleotides at each CpG site averaged by the number of CpG sites. Statistical differences between controls, cell lines, or tissues and healthy WBCs were computed using the Mann–Whitney *U* test ($P_{Fully\_meth.}$ = 9.97e−07, $P_{Unmeth.}$ = 1.86e−06, P_{CRC_Cells} = 0.266, P_{OVC_Cells} = 1.20e−02, P_{BRC_Cells} = 2.47e−06, P_{UVM_Cells} = 6.77e−05, $P_{Healthy\_OVC\_Tissues}$ = 0.063, $P_{OVC\_Tissues}$ = 8e−03, $P_{BRC\_Tissues}$ = 4.10e−04, and $P_{UVM\_Tissues}$ = 0.88; Supplementary Table S3). **B,** Global DNA methylation in cancer plasma including metastatic stages (M+) and nonmetastatic stages (M0) as well as HD plasmas. Statistical differences between each cancer subgroup and healthy samples were computed using the Mann–Whitney *U* test (P_{CRC_M+} = 1.27e−29, P_{BRC_M+} = 3.79e−19, P_{UVM_M+} = 8.29e−06, P_{LC_M+} = 0.655, P_{OVC_M0} = 1.94e−05, P_{GAC_M0} = 4.28e−08, and P_{BRC_M0} = 9.10e−01; Supplementary Table S3). Black dotted lines represent the median. **C,** Methylation level at each targeted CpG site (*x*-axis), for each healthy sample (*y*-axis), depicted as a heatmap. CpG numbers are indicated. The metaplot represents the average methylation levels of the population. Amplicon numbers are indicated. **D,** Differential methylation levels between healthy samples and patients for each type of cancer, represented as metaplots. **E,** Proportion of methylation motifs, called haplotypes, for each amplicon (mean centered per amplicon). Only the most important features are represented (see **Fig. 3F**; "Materials and Methods"). Blue arrows highlight the most abundant haplotype in each amplicon. **F,** Mean centered abundance of the most important haplotypes with the highest co-methylation patterns (mostly fully methylated or fully unmethylated molecules) in cancer subgroups compared with HDs. Statistical significances were computed using the Mann–Whitney *U* test on raw haplotype proportions (Supplementary Table S5). HD, healthy donor; WBC, white blood cell.

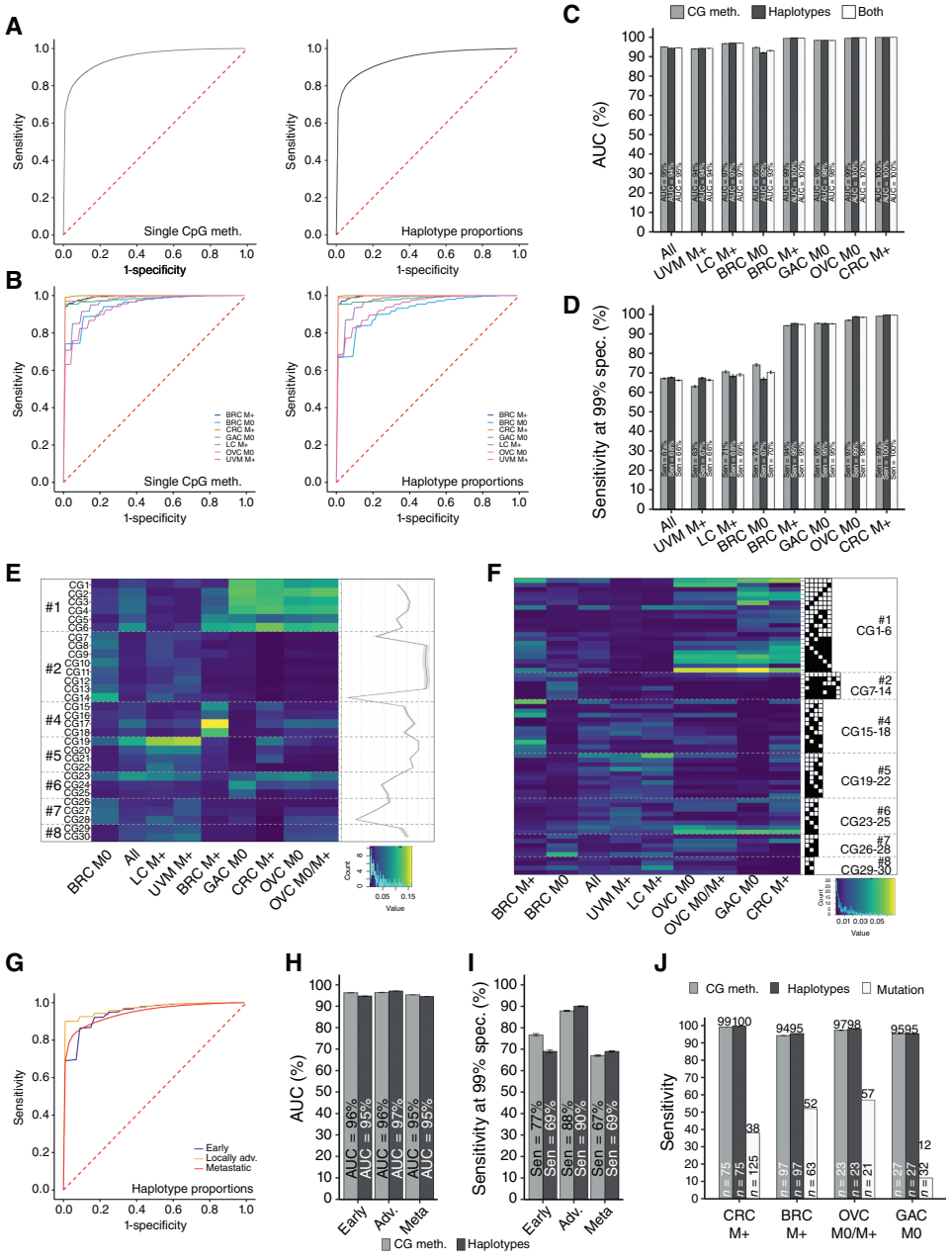

Figure 3.

L1PA hypomethylation–based classifiers recognize samples from multiple forms of cancer. **A** and **B,** ROC curves obtained for plasma sample classification using single CpG methylation levels ($n = 30$) or haplotype proportions ($n = 372$) with the "all cancers" model (**A**) or the "cancer-types" models (**B**). All classifications include 5,000 stratified random repetitions of learning on 60% of the samples and testing on the 40% left, with undersampling for classes equilibrium (results with and without undersampling are presented in Supplementary Fig. S3A and S3B). $N_{CRC\_M+} = 75$, $N_{BRC\_M+} = 97$, $N_{LC\_M+} = 50$, $N_{UVM\_M+} = 55$, $N_{OVC\_M+} = 4$ (included only in "all cancers" testing), $N_{OVC\_M0} = 18$, $N_{GAC\_M0} = 27$, and $N_{BRC\_M0} = 23$ tested vs. 123 HDs. ROC curves shown are obtained by averaging the sensitivity and specificity of each repetition of learning. **C** and **D,** Performances for classifiers using single CpG methylation levels (gray), haplotype proportions (black), or both (white) presented as AUCs (**C**) or sensitivities at 99% specificity (**D**). Average AUCs are computed from the 5,000 AUCs generated by each repetition of learning. Bars indicate 95% CI. **E** and **F,** Importance (mean decrease in impurity) of the features used by the classifiers depicted as clustered heatmaps. The features correspond to the CpG targets (**E**) or the haplotypes (**F**). Only the most important haplotypes (feature importance level >1%) are shown. **G,** ROC curves obtained for plasma sample classification with the three-stage model, using haplotype features. **H** and **I,** Performances for the three-stage classifiers using single CpG methylation levels (gray) or haplotype proportions (black) presented as AUCs (**H**) or sensitivities at 99% specificity (**I**). Early stages (I/II, $N = 31$), locally advanced stages (III, $N = 30$), and metastatic stages (IV, $N = 281$). **J,** Cancer detection rates with the methylation-based DIAMOND assay (haplotypes and CG methylation) vs. common recurrent mutations for samples assessed in previous studies [(13, 14, 62, 63)] or with NGS (Supplementary Table S6). HD, healthy donor. NGS, next-generation sequencing.

patterns in HR^+ $HER2^-$ BRC versus TNBC (Supplementary Fig. S6F–S6H). We observed that haplotype features showed consistent patterns with important CG features (**Fig. 3F**). However, haplotypes provide a more detailed view of the methylation patterns with a strong importance of the most methylated or nonmethylated molecules. We still observed that some methylation intermediates are important for cancer detection (ex: in amplicon #1 in CRC M+ and GAC M0, #2 in BRC M0, #4 in BRC M+ and other subgroups, #5 in LC M+ and UVM M+, #7 in OVC, and #8 in BRC M0 and LC M+). BRC M+ and M0 subgroups cluster together, showing that even if they are not highly similar, they are closer to each other than to other cancer types. These breast cancer specificities are worth exploring further in the future. Overall, this suggests that L1PA methylation alterations detected from cfDNA vary in different types and stages of cancer.

To estimate the ability of DIAMOND to detect cancer at early stages of the disease, we build classifiers for three stage classes gathering all cancer types: early stages (I/II, $N = 31$), locally advanced stages (III, $N = 30$), and metastatic stages (IV, $N = 281$). Classifications were highly performant for all three stage categories ($AUC_{Early} = 95\%$, $AUC_{Adv.} = 97\%$, and $AUC_{Meta} = 95\%$; **Fig. 3G–H**; Supplementary Fig. S6I and S6D) with a mean sensitivity of 70% for early stages, ($Sen_{Early} = 70\%$, $Sen_{Adv.} = 90\%$, and $Sen_{Meta} = 69\%$; **Fig. 3I**; Supplementary Fig. S6E).

Next, we analyzed WGBS from healthy and cancer plasma from two recently published studies (55, 56) to evaluate if we can retrieve L1PA hypomethylation signal with non-targeted methods (see "Materials and Methods"; Supplementary Fig. S7A). Consistent with our findings, we observed statistically significant hypomethylation in cancer samples (Supplementary Fig. S7B–S7H). These whole genome approaches were nevertheless not as efficient as the DIAMOND-targeted approach to automatically detect cancer from L1PA hypomethylation as they cover less well the regions of interest (Supplementary Fig. S7A, S7F, and S7I). Indeed, they involve mapping on a reference genome leading to loss of data. Moreover, L1PA methylation measured with DIAMOND largely outperforms methods based on the detection of mutations. In comparison, the identification of the same tumor samples via the detection of frequent recurrent mutations, which is commonly used in the clinic, does not exceed 57% for ovarian cancer (Supplementary Table S8), 38% for colon cancer (62), and 52% for metastatic breast cancer (**Fig. 3J**; refs. 13, 14). We particularly achieved remarkable performance on the cohort of 27 localized gastric cancers with a detection rate of 95% of true positive as compared with 12% for mutation screening (63). This is mostly due to the fact that methylation changes occur in virtually all patients with cancer, unlike recurrent mutations.

Multicancer classification performances are reproducible on an independent cohort

To validate the DIAMOND approach, we tested a second independent cohort consisting of 214 patients affected with the same types of cancers as in the first cohort, excluding uveal melanoma and nonmetastatic gastric cancers, along with 60 healthy donors (**Fig. 4A**). First, we confirmed that the methylation patterns along the L1 structure were highly reproducible between healthy donors from cohorts 1 and 2, at the level of single CpG targets (**Fig. 4B**) but also for haplotype proportions (**Fig. 4C**; Supplementary Table S9). Although methylation at single CpG within cancer subgroups showed slightly more variability (Supplementary Fig. S8A), global methylation levels were quite reproducible between the two cohorts, showing similar distributions and no statistical differences (**Fig. 4D**; Supplementary Table S10), except for nonmetastatic ovarian cancers. There was an important heterogeneity among the OVC M0 samples of cohort 2, which clustered into two distinct groups, whereas cohort 1 was more homogeneous (Supplementary Fig. S8B). Notably, no correlation was found with available clinicohistopathologic parameters (age, staging, CA125 level, mutational status, treatment, or response to therapy). Differential haplotype proportions between healthy and cancer subgroups were also mostly conserved (Supplementary Fig. S8C; Supplementary Table S11). Overall, the method showed good reliability with the 7-amplicon panel used and good robustness in detecting L1 methylation levels and changes.

Because age-related changes in DNA methylation have been described (64, 65) and that the healthy donors included in the study are younger overall than the patients with cancer (Supplementary Fig. S9A), we have investigated whether there was an effect on the methylation patterns we studied. We found a significant but very small effect which seemed much smaller that the effect of disease status (Supplementary Fig. S9B and S9C). This small effect was tending toward an increase in methylation with age (Supplementary Fig. S9D), and we observed similar patterns and differences between healthy and cancer samples when adjusting for the age (Supplementary Fig. S9E–S9G; Supplementary Table S12). Furthermore, we observed similar performances in age-matched and non–age-matched cohorts extracted from C2 (Supplementary Fig. S9H–S9J), demonstrating that age is not a confounding factor.

To validate our classifiers, we trained models on the entire first cohort and evaluated them on the second set of independent samples. The results showed excellent classification performances with an overall AUC of 88% when testing all cancers together with no annotations of their histologic types and AUC between 88% and 100% for the cancer subgroup "expert" models with sensitivities at 99% specificity between 49% and 100% (**Fig. 4E**), including 55% for localized breast cancers. It was, however, lower for metastatic lung cancer with a mean sensitivity of 49%. We observed that haplotype models were more robust compared with single CpG methylation rates (Supplementary Fig. S8D–S8G). This could be explained by the fact that haplotypes capture true methylation patterns at the molecule level, enabling to discard noise caused by experimental variability for example. Next, we applied the same validation method, training on C1 and testing on C2, for the three-stage "expert" classifiers and observed great classification performances with a mean AUC of 99% and a mean sensitivity of 79% for early stages (**Fig. 4F**). When comparing the performances for each subgroup with the classification rates extracted from the model "all," we observed similar AUCs for most cancers, but lower rates for the BRC M0 subgroup and the early-stage group (Supplementary Fig. S10A), which is mostly composed of BRC M0 (Supplementary Fig. S10B).

To further investigate the generalization of our marker, we have also tested the ability of DIAMOND to detect cancers of a type or subgroup for which the model is not trained. To do so, we have combined the data from cohorts C1 and C2 and trained on the whole data set, systematically removing one cancer subgroup or one cancer type and testing specifically for this subgroup or type (see "Materials and Methods"; **Fig. 4G**; Supplementary Fig. S10C). We observed performances similar to when the subgroup is included in the train set, which demonstrates the universality of L1PA hypomethylation. Classification performance is noticeably lower for the BRC M0 subgroup when the model still performs well for OVC M0 and GAC M0. This may be because the BRC M0 subgroup

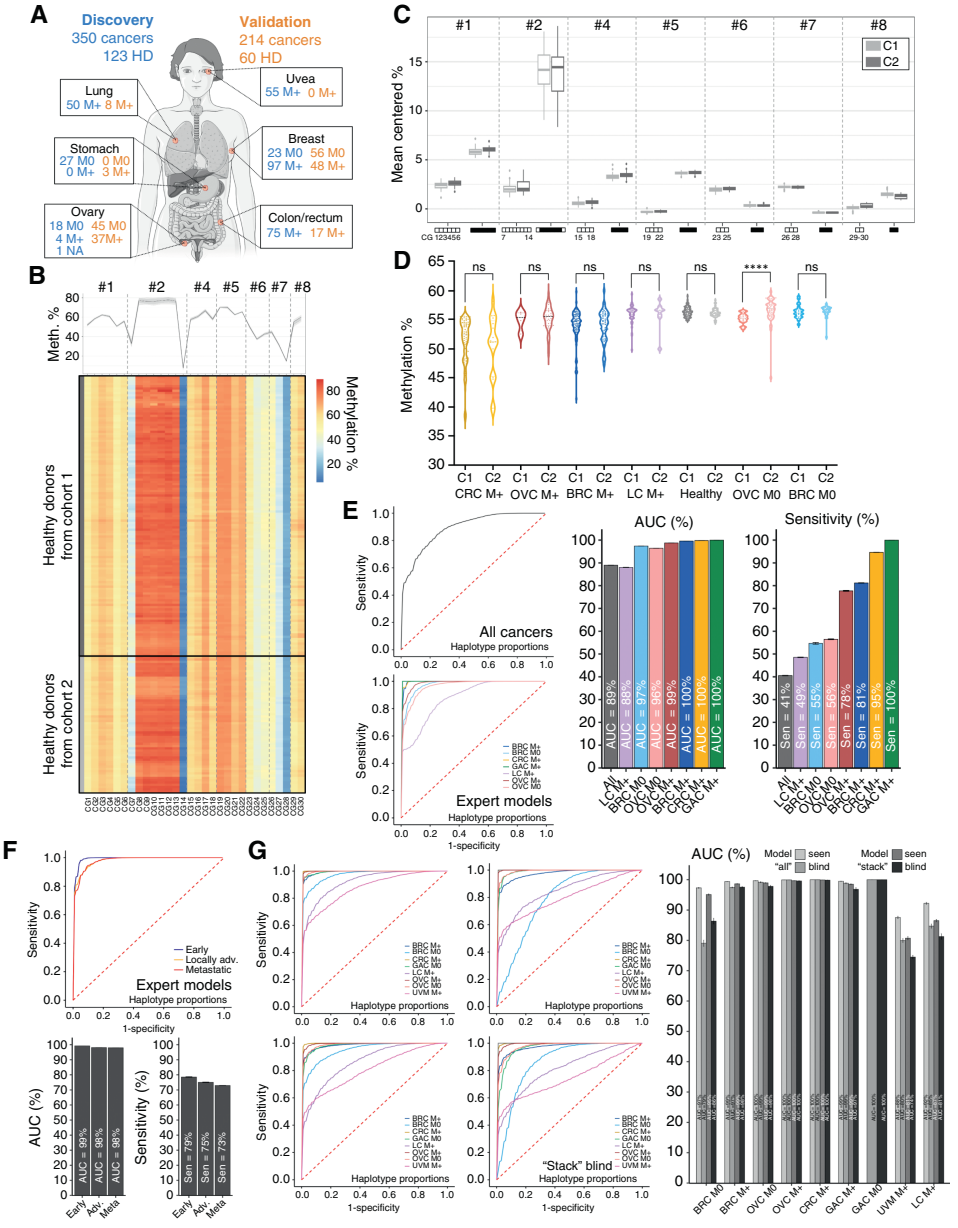

Figure 4.

Multicancer classification performances are reproducible on an independent cohort. **A,** Number of patients and HDs in the discovery cohort (C1) and in the validation cohort (C2) for each cancer type and dissemination stage (nonmetastatic: M0 vs. metastatic: M+, NA, stage not available). Generated using Servier Medical Art. **B,** Methylation level at each targeted CpG sites (*x*-axis), for each healthy sample (*y*-axis) from C1 vs. C2, depicted as a heatmap. No clustering is done on the data, which come ordered by targeted CpG site on the *x*-axis (amplicon numbers are indicated). The metaplots represent the average levels for donors of C1 vs. C2 at each CpG site. **C,** Mean centered abundance of the most important haplotypes, with the highest co-methylation patterns, in HDs from C1 vs. C2. (Statistical differences computed using the Mann–Whitney *U* test are available in Supplementary Table S9) **D,** Comparison of the global levels of methylation in C1 vs. C2. Methylation levels are calculated as explained previously in **Fig. 2**. The *P* values are computed using the Mann–Whitney *U* test ($P_{CRC\_M+} = 0.680$, $P_{OVC\_M+} = 0.816$, $P_{BRC\_M+} = 0.783$, $P_{LC\_M+} = 0.596$, $P_{Healthy} = 0.316$, $P_{OVC\_M0} = 4.74e{-}05$, $P_{BRC\_M0} = 0.132$; Supplementary Table S10). Black dotted lines represent the median. **E,** Performances for validation classifiers using haplotype features presented as ROC curves, AUCs, and sensitivities at 99% specificity obtained with the "all" cancers model or the "expert" models for cancer subgroups. All classifications include 5,000 stratified random repetitions of learning on the whole discovery cohort and testing on the whole validation cohort without undersampling. ROC curves shown are obtained by averaging the sensitivity and specificity of each repetition of learning. Average AUCs are computed from the 5,000 AUCs generated by each repetition of learning. Bars indicate 95% CI. **F,** Performances for three-stage "expert" classifiers: early stages (I/II, $N_{C1} = 31$, $N_{C2} = 38$), locally advanced stages (III, $N_{C1} = 30$, $N_{C2} = 54$), and metastatic stages (IV, $N_{C1} = 281$, $N_{C2} = 113$), presented as mean ROC curves, AUCs, or sensitivities at 99% specificity. **G,** Performances for integrated models ("all" or "stack") when training for the specific group tested (seen) vs. when not training for this subgroup (blind). These classifications have been performed on the whole sample set, including C1 and C2. HD, healthy donor.

contains the earliest stages of both cohorts (Supplementary Fig. S10B) and the model is less well trained for these stages when removing it. To improve this, we have developed another integrated model that is trained for all cancer types and stages together but also incorporates the probabilities from the "expert" models, defining profiles for various stages and dissemination status. This "stack" model provides a unique prediction for cancer versus healthy status for each sample at once, independently of their cancer type/dissemination status, and allows to balance for the cancer characteristics used to train the model by including them all. We have then extracted the performances per cancer subgroups out of this integrated model, which reached better performances to classify BRC M0 and early-stage samples (see "Materials and Methods"; **Fig. 4G**; Supplementary Fig. S10A).

This demonstrates the robustness of cancer detection by probing L1PA hypomethylation from plasma DNA with the DIAMOND assay and its ability to detect early stages. Finally, we evaluated if the extent of L1PA hypomethylation was associated with survival. In the validation cohort, we compared patients with high methylation to those with low methylation levels (above or below the median threshold—Supplementary Fig. S10D) and observed that more pronounced hypomethylation is clearly associated with shorter survival (Supplementary Fig. S10E). This effect is thought to mainly reflect the tumor burden, which directly impacts the fraction of tumor DNA circulating in the blood, but it remains an interesting noninvasive tumor marker.

DIAMOND data contain signal to infer the tumor burden, which improves cancer detection

When looking at the overall methylation rates, we detected significantly more hypomethylation for more advanced stages of the disease, in particular in metastatic stages compared with localized stages (**Fig. 5A**; Supplementary Table S13). However, there was no significant differences between metastatic tumor tissues and primary tissues (**Fig. 5B**; Supplementary Table S14), which confirmed that L1PA methylation alteration is an early event in carcinogenesis (66, 67) and also affects early-stage cancers. The differences observed in the blood (**Fig. 5A**) reflect the ctDNA fraction, which is known to be directly influenced by the tumor burden and the stage of the disease (1, 9). This demonstrates the quantitative potential of DIAMOND, which could help quantify the tumor burden and monitor the disease.

A recognized marker to noninvasively estimate the tumor burden and the fraction of ctDNA is the aneuploidy or CNAs (68, 69), a hallmark of cancer genomes (70). Given that DIAMOND hits are dispersed throughout the genome (**Fig. 1B**), we investigated the possibility to use our data to perform CNA analysis. The mFast-SeqS approach had previously used a PCR-based L1PA targeting as a prescreening tool to estimate the fraction of ctDNA (71, 72). This was done on native DNA whereas our data resulted from bisulfite-treated DNA.

We first tested this approach on 15 breast cancer cell lines that were also assessed by CytoScan HD microarrays for aneuploidy. DIAMOND provided an average of 78,000 uniquely mappable reads per cell line, corresponding to around 10,000 L1PA copies precisely located in the genome. These L1PA hits homogeneously overlapped with regions covered by CytoScan probes along the genome (**Fig. 5C**). We computed *z*-scores, quantifying CNAs, at the level of chromosome arms as previously described (see "Materials and Methods"; refs. 57, 72) and obtained similar results to those found with CytoScan arrays (Supplementary Fig. S11A). We observed low alteration scores for the normal-like breast cell line HTERT-HME1 and good correlations between the two methods for the majority of the cell lines (Supplementary Fig. S11B).

Next, we computed genome-wide *z*-scores in healthy and cancer plasma samples and observed high alteration scores specifically in cancer samples (**Fig. 5D**). Cancer subgroup *z*-scores mirrored global hypomethylation profiles (**Figs. 2B**, **4D**, and **5E**; Supplementary Table S15), both reflecting tumor burden and ctDNA fractions available. However, global methylation rates and *z*-scores were only moderately anticorrelated (**Fig. 5F**), demonstrating that these are partially independent markers that can provide distinct signals (Supplementary Fig. S11C).

To obtain a final classification labeling each sample as healthy or cancer, we used a two-step categorization incorporating CNA analysis, which improved cancer detection. We used the probability of the cancer prediction provided by the methylation-based validation models ("expert" and "all" models), applying a threshold identified on the discovery cohort, followed by a reclassification of samples which presented a *z*-score >121, as cancer. This cut-off value was deduced from a cross-validation applied on C1 (see "Materials and Methods"; Supplementary Fig. S11C). This classifier achieved high sensitivities with specificities between 97% and 100% for the "expert" models and 93% for the integrated "all" model in four distinct cancer types (breast cancer, colorectal cancer, lung cancer, and ovarian cancer; **Fig. 5G**) and could be applied as is in the clinic. This was particularly promising for localized breast cancer with a sensitivity of 100% with both model types.

Discussion

In this study, we established a robust proof of concept that targeting hypomethylation of retrotransposons from cfDNA is a sensitive and specific biomarker to detect multiple forms of cancer noninvasively. Repetitive regions provide genome-wide information as they hold half of the CpG sites present in the human genome (73). Hypomethylation of L1 elements, which is a common feature of multiple forms of cancer, is a universal marker and helps cover the heterogeneous profiles of patients with cancer in a single test. Previous methylation studies have left these regions aside as they are inherently difficult to map, and differentially methylated region analysis is commonly performed on mapped data. However, repeats were previously used to profile aneuploidy and one of the first reported liquid biopsy cancer detection test was based on serum DNA integrity calculated with the ratio of short over long Alu fragments (74).

Here, we developed a new pipeline to detect methylation profiles at repeats with a single bp resolution, without resorting to mapping on a reference genome. This allowed us to rescue unmappable sequences (such as polymorphic copies that are not annotated in the reference genomes) or sequences that are difficult to map (such as multi-mappers) and retain most of our data. This is instrumental in achieving high sensitivity. The DIAMOND assay demonstrated high performance in detecting cancer samples and we established its feasibility in six different cancer types, including three at localized stages. It outperforms mutation screening, as it covers virtually all patients and presents a promising marker for pan-cancer detection.

Previous cfDNA multicancer tests enabled sensitive detection and classification of multiple forms of cancers such as the CancerSeek test, based on the combination of mutations and proteins detection (11, 75) and more recently combined with aneuploidy profiling using representation of Alu sequences (76) or the cfDNA methylation test

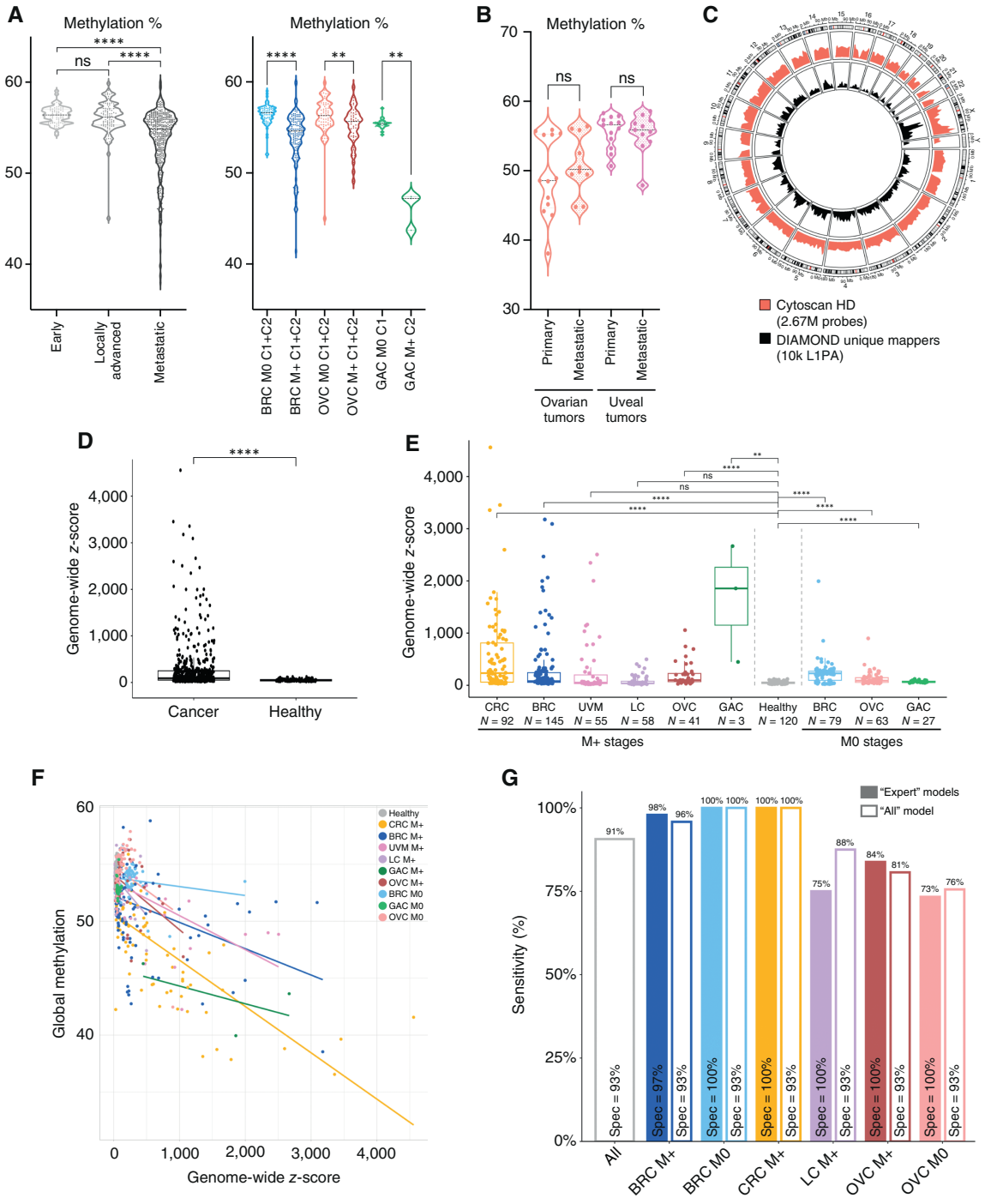

Figure 5.

DIAMOND data contain signal to infer the tumor burden, which improves cancer detection. **A** and **B,** Comparison of the average levels of methylation observed in localized vs. metastatic plasma samples (**A**, $P_{Early/Adv.} = 0.327$, $P_{Adv./Meta.} = 3.14e{-}11$, $P_{Early/Meta.} = 2.82e{-}14$; $P_{BRC\_M0/M+} = 1.3e{-}18$, $P_{OVC\_M0/M+} = 0.006$, and $P_{GAC\_M0/M+} = 0.005$; Supplementary Table S13) or in primary vs. metastatic tissues (**B**, $P_{OVC} = 0.257$; $P_{UVM} = 0.820$; Supplementary Table S14). **C,** L1PA unique hits obtained for 15 breast cancer cell lines compared with the distribution of CytoScan probes distributed throughout the human genome. **D,** Genome-wide *z*-score for all cancer ($N = 564$) vs. healthy plasma samples ($N = 120$, 63 of the total 183 HDs are used as references to compute the *z*-score and are not displayed here, $p = 1.21e{-}20$). **E,** Genome-wide *z*-score by cancer subgroups vs. healthy samples. The *P* values are computed using the Mann–Whitney *U* test ($P_{CRC\_M+} = 2.05e{-}18$, $P_{BRC\_M+} = 1.01e{-}18$, $P_{UVM\_M+\_} = 0.169$, $P_{LC\_M+} = 0.769$, $P_{OVC\_M+\_} = 1.84e{-}11$, $P_{GAC\_M+} = 0.003$, $P_{BRC\_M0} = 5.12e{-}17$, $P_{OVC\_M0} = 1.09e{-}12$, and $P_{GAC\_M0} = 8.40e{-}06$; Supplementary Table S15). **F,** Correlation analysis for genome-wide *z*-score vs. global methylation ($r_{overall} = -0.62$; $P = 1.25e{-}69$). **G,** Performances of the two-step model incorporating CNA with DNA methylation analysis (classification is done as follows: $Proba_{Cancer} \leq$ Threshold C1 AND GZ-score ≤ 121: prediction = Healthy; $Proba_{Cancer} >$ Threshold C1 OR GZ-score > 121: prediction = Cancer, see "Materials and Methods"). HD, healthy donor.

Galleri from GRAIL, which targets 1,100,000 CpGs within unique regions (39, 40). In comparison, DIAMOND targets about 100,000 CpG sites within 30 to 40,000 copies of LINE-1 using a unique set of probes for all these copies and the various cancer type tested. Very recently, several multicancer detection methods based on retrotransposons were published. This includes the detection of the circulating L1 protein ORF1p (67) as well as the probing of the representation of all type of repeats from Annapragada and colleagues (77). Another study demonstrated that cfDNA cleavage profiles at Alu sequences reflect their methylation status and can be used to detect liver cancer and nasopharyngeal carcinoma (78).

The DIAMOND assay interrogates specifically the methylation status of L1 to obtain a global representation of the hypomethylation occurring during carcinogenesis and assess the potential of circulating DNA methylation changes at L1PA elements as a universal tumor biomarker. Our aim was to develop a new highly sensitive strategy to detect cancer-specific signatures in blood. Overall, DIAMOND reached great performances for multicancer detection using "expert" models for specific cancer subgroups or integrated models including all types of cancers. Lower detection rate in metastatic lung cancer may be related to the fact that these samples seem to have a low tumor burden as indicated by their genome-wide *z*-scores (**Fig. 5E**). However, integrating other regions with cancer-specific methylation changes could help improve detecting this type of cancer.

The DIAMOND assay provides methylation profiles from minute amount of cfDNA, down to a few nanograms, with high precision and high coverage using an affordable sequencing depth. Stronger L1PA hypomethylation was associated with shorter overall survival demonstrating its prognostic value. We therefore anticipate that our method has the potential to be applied for the development of routine clinical tests. Although our integrated models provide proof of concept for cancer detection with blood screening in asymptomatic patients, the "expert" models could also be useful to help diagnosis when we suspect the cancer location or to perform disease follow up.

To push the DIAMOND assay toward a clinically applicable test, we also demonstrated that DIAMOND data can be used to perform CNAs analysis which improves cancer detection. We integrated this analysis in a classifier providing "healthy" or "cancer" labels for each sample and reached a detection of 91% of true positives for all cancers together and in particular a 100% sensitivity with 100% specificity for localized breast cancer with the BRC M0 expert model.

Further testing with a larger number of samples covering earlier stages, including an independent cohort of gastric nonmetastatic cancers, more subtypes and different types of cancer will enable to consolidate and expand these findings. Moreover, this will strengthen the classification models, which will perform better with more samples for training and testing. In the current study, we did not control for ethnic origin of samples tested. The potential impact of L1 polymorphism linked to ancestry should be estimated in future studies. It will also be important to study the impact of other conditions, such as autoimmune diseases, which may lead to the detection of L1 hypomethylation in blood. The recent study on the detection of circulating L1 ORF1p in cancer by Taylor and colleagues (67) demonstrated a high specificity and no sign of L1 reactivation in blood of patients with autoimmune disease, indicating that it might be a cancer-specific phenomenon. DIAMOND analysis could further be used to infer the tumor burden and monitor the disease to better detect minimal residual disease and the relapse early. However, the impact of treatments on methylation status should be investigated first.

Overall, we developed a *turnkey* analysis method that identifies tumor plasmas across multiple types of cancer with the same marker. This approach offers an optimized balance between the number of targeted regions and sequencing depth, which could extensively improve the sensitivity of ctDNA detection in a cost-effective manner and improve management of patients with cancer.

Authors' Disclosures

M. Michel reports a patent for PCT/EP2023/074092—Sensitive and Specific Determination of DNA Methylation Profiles pending. M. Heidary reports grants from the European Research Council (ERC-StG EpiDetect), the Ligue contre le cancer (RS17-75-75), the prematuration program of the Centre National pour la Recherche Scientifique, the SiRIC 2 Curie program (INCa-DGOS-Inserm_12554), the DEEP Strive funding (LABEX DEEP 11-LBX0044), ANR-10-EQPX-03 (Equipex), ANR-10-INBS-09-08 (France Genomique Consortium), and ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) during the conduct of the study, as well as a patent for "PCT/EP2023/074092—Sensitive and Specific Determination of DNA Methylation Profiles—Inventors: Proudhon, Charlotte; Azencott, Chloé-Agathe; Michel, Marc; Heidary, Maryam" pending. M. Kamal reports personal fees from Roche outside the submitted work. C. Le Tourneau reports personal fees from MSD, Bristol Myers Squibb, Merck, AstraZeneca, Celgene, Seattle Genetics, Roche, Novartis, Rakuten, Nanobiotix, and GSK outside the submitted work. M.-H. Stern reports a patent for US20190256921A1 pending, a patent for ES2978017T3 issued, a patent for PCT/EP2019/056445 issued, and a patent for PCT/EP2023/057543 issued. C.-A. Azencott reports grants from Agence Nationale de la Recherche during the conduct of the study, as well as grants from Janssen Research & Development outside the submitted work; in addition, C.-A. Azencott reports a patent for WO2024047250—Sensitive and Specific Determination of DNA Methylation Profiles pending. C. Proudhon reports grants from the European Research Council, the Ligue contre le cancer, the French National Center for Scientific Research, grants from SiRIC 2 Curie program, and the DEEP Strive funding during the conduct of the study, as well as a patent for PCT/EP2023/074092—Sensitive and Specific Determination of DNA Methylation Profiles—Inventors: Proudhon, Charlotte; Azencott, Chloé-Agathe; Michel, Marc; Heidary, Maryam pending. No disclosures were reported by the other authors.

Authors' Contributions

M. Michel: Conceptualization, data curation, software, formal analysis, investigation, visualization, methodology, writing–original draft. **M. Heidary:** Conceptualization, data curation, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **A. Mechri:** Data curation, formal analysis, validation, investigation, visualization, writing–review and editing. **K. Da Silva:** Data curation, software, formal analysis, visualization, writing–review and editing. **M. Gorse:** Data curation, software, formal analysis, validation, investigation, visualization, writing–review and editing. **V. Dixon:** Data curation, investigation, visualization, writing–review and editing. **K. von Grafenstein:** Data curation, software, formal analysis, validation, visualization, writing–review and editing. **C. Bianchi:** Data curation, investigation, visualization, writing–review and editing. **C. Hego:** Investigation. **A. Rampanou:** Investigation. **C. Lamy:** Resources, data curation. **M. Kamal:** Resources, data curation. **C. Le Tourneau:** Resources, data curation. **M. Séné:** Investigation. **I. Bièche:** Resources, writing–review and editing. **C. Reyes:** Investigation. **D. Gentien:** Resources, investigation. **M.-H. Stern:** Resources. **O. Lantz:** Resources. **L. Cabel:** Resources. **J.-Y. Pierga:** Resources, writing–original draft. **F.-C. Bidard:** Resources, writing–original draft. **C.-A. Azencott:** Software, formal analysis, supervision, validation, writing–original draft, writing–review and editing. **C. Proudhon:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing.

Acknowledgments

We thank the members of C. Proudhon's laboratory for critical reading of the manuscript. We are grateful to D. Bourc'his and her team for hosting us during

part of this study. We thank the members of the ICGex next-generation sequencing platform of the Institut Curie, especially S. Lameiras, V. Raynal, and S. Baulande for advice and the nonprofit organization "La Vannetaise" for financial support. The next-generation sequencing facility was supported by ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) grants and by the Cancéropôle Île-de-France. This research was supported by grants, of which C. Proudhon was recipient, from the European Research Council (ERC-StG EpiDetect), the Ligue contre le cancer (RS17-75-75), the prematuration program of the Centre National pour la Recherche Scientifique, the SiRIC 2 Curie program (INCa-DGOS-Inserm_12554), and the DEEP Strive funding (LABEX DEEP 11-LBX0044). C.-A. Azencott research was supported in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Note

Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Received August 16, 2024; revised September 20, 2024; accepted November 22, 2024; published first December 2, 2024.

References

1. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra24.
2. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med 2014;20:548–54.
3. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med 2015;7:302ra133.
4. Bidard F-C, Weigelt B, Reis-Filho JS. Going with the flow: from circulating tumor cells to DNA. Sci Transl Med 2013;5:207ps14.
5. Diaz LA, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol 2014;32:579–86.
6. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic—implementation issues and future challenges. Nat Rev Clin Oncol 2021;18:297–312.
7. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet 2019;20:71–88.
8. Stover DG, Parsons HA, Ha G, Freeman SS, Barry WT, Guo H, et al. Association of cell-free DNA tumor fraction and somatic copy number alterations with survival in metastatic triple-negative breast cancer. J Clin Oncol 2018;36:543–53.
9. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 2017;17:223–38.
10. Mattox AK, Bettegowda C, Zhou S, Papadopoulos N, Kinzler KW, Vogelstein B. Applications of liquid biopsies for cancer. Sci Transl Med 2019;11:eaay1984.
11. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359:926–30.
12. Riva F, Bidard F-C, Houy A, Saliou A, Madic J, Rampanou A, et al. Patient-specific circulating tumor DNA detection during neoadjuvant chemotherapy in triple-negative breast cancer. Clin Chem 2017;63:691–9.
13. Jeannot E, Darrigues L, Michel M, Stern M-H, Pierga J-Y, Rampanou A, et al. A single droplet digital PCR for ESR1 activating mutations detection in plasma. Oncogene 2020;73:2987–95.
14. Darrigues L, Pierga J-Y, Bernard-Tessier A, Bieche I, Silveira AB, Michel M, et al. Circulating tumor DNA as a dynamic biomarker of response to palbociclib and fulvestrant in metastatic breast cancer patients. Breast Cancer Res 2021;23:31.
15. Silveira AB, Bidard F-C, Tanguy M-L, Girard E, Trédan O, Dubot C, et al. Multimodal liquid biopsy for early monitoring and outcome prediction of chemotherapy in metastatic breast cancer. NPJ Breast Cancer 2021:7;115.
16. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer 2011;11:726–34.
17. Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. Science 2017;357:eaal2380.
18. van der Pol Y, Mouliere F. Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. Cancer Cell 2019;36:350–68.
19. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 2007;8:286–98.
20. Feinberg AP. The key role of epigenetics in human disease prevention and mitigation. N Engl J Med 2018;378:1323–34.
21. Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet 2017;49:635–42.
22. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res 2018;46:e89–9.
23. Moarii M, Reyal F, Vert J-P. Integrative DNA methylation and gene expression analysis to assess the universality of the CpG island methylator phenotype. Hum Genomics 2015;9:26.
24. Chan KCA, Jiang P, Chan CWM, Sun K, Wong J, Hui EP, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci U S A 2013;110:18761–8.
25. Legendre C, Gooden GC, Johnson K, Martinez RA, Liang WS, Salhia B. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. Clin Epigenetics 2015;7:100.
26. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A 2015;112:E5503–12.
27. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A 2021;113:E1826–34.
28. Garrigou S, Perkins G, Garlan F, Normand C, Didelot A, Le Corre D, et al. A study of hypermethylated circulating tumor DNA as a universal colorectal cancer biomarker. Clin Chem 2016;62:1041–3.
29. Barault L, Amatu A, Siravegna G, Ponzetti A, Moran S, Cassingena A, et al. Discovery of methylated circulating DNA biomarkers for comprehensive noninvasive monitoring of treatment response in metastatic colorectal cancer. Gut 2018;67:1995–2005.
30. Jin S, Zhu D, Shao F, Chen S, Guo Y, Li K, et al. Efficient detection and post-surgical monitoring of colon cancer with a multi-marker DNA methylation liquid biopsy. Proc Natl Acad Sci U S A 2021;118:e2017421118.
31. Lun FMF, Chiu RWK, Sun K, Leung TY, Jiang P, Chan KCA, et al. Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. Clin Chem 2013;59:1583–94.
32. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nat 2018;563:579–83.
33. Nassiri F, Chakravarthy A, Feng S, Shen SY, Nejad R, Zuccato JA, et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. Nat Med 2020;26:1044–7.
34. Nuzzo PV, Berchuck JE, Korthauer K, Spisak S, Nassar AH, Abou Alaiwi S, et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. Nat Med 2020;26:1041–3.
35. Liu X, Ren J, Luo N, Guo H, Zheng Y, Li J, et al. Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by methylated CpG tandem amplification and sequencing (MCTA-Seq). Clin Epigenetics 2019;11:93.
36. Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Sci Transl Med 2020;12:eaax7533.

37. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun 2020;11:3475.
38. Cao F, Wei A, Hu X, He Y, Zhang J, Xia L, et al. Integrated epigenetic biomarkers in circulating cell-free DNA as a robust classifier for pancreatic cancer. Clin Epigenetics 2020;12:112.
39. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV; CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol 2020;31:745–59.
40. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol 2021;32:1167–77.
41. Ross JP, Rand KN, Molloy PL. Hypomethylation of repeated DNA sequences in cancer. Epigenomics 2010;2:245–69.
42. Burns KH. Transposable elements in cancer. Nat Rev Cancer 2017;17:415–24.
43. Rodić N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. Am J Pathol 2014;184:1280–6.
44. Lanciano S, Philippe C, Sarkar A, Pratella D, Domrane C, Doucet AJ, et al. Locus-level L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites. Cell Genom 2024;4:100498.
45. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet 2020;52:306–19.
46. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res 2016;26:745–55.
47. Gainetdinov IV, Kapitskaya KY, Rykova EY, Ponomaryova AA, Cherdyntseva NV, Vlassov VV, et al. Hypomethylation of human-specific family of LINE-1 retrotransposons in circulating DNA of lung cancer patients. Lung Cancer 2016;99:127–30.
48. Nagai Y, Sunami E, Yamamoto Y, Hata K, Okada S, Murono K, et al. LINE-1 hypomethylation status of circulating cell-free DNA in plasma as a biomarker for colorectal cancer. Oncotarget 2017;8:11906–16.
49. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res 2021;31:1280–9.
50. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 2016;164:57–68.
51. Breiman L. Random forests. Mach Learn 2001;45:5–32.
52. Boulesteix A, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov; 2012;2:493–507.
53. Barandela R, Sánchez JS, García V, Rangel E. Strategies for learning in class imbalance problems. Pattern Recogn 2003;36:849–51.
54. Basse C, Morel C, Alt M, Sablin MP, Franck C, Pierron G, et al. Relevance of a molecular tumour board (MTB) for patients' enrolment in clinical trials: experience of the Institut Curie. ESMO Open 2018;3:e000339.
55. Liu Y, Reed SC, Lo C, Choudhury AD, Parsons HA, Stover DG, et al. FinaleMe: predicting DNA methylation by the fragmentation patterns of plasma cell-free DNA. Nat Commun 2024;15:2790.
56. Gao Y, Zhao H, An K, Liu Z, Hai L, Li R, et al. Whole-genome bisulfite sequencing analysis of circulating tumour DNA for the detection and molecular classification of cancer. Clin Transl Med 2022;12:e1014.
57. Belic J, Koch M, Ulz P, Auer M, Gerhalter T, Mohan S, et al. Rapid identification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. Clin Chem 2015;61:838–49.
58. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, et al. Statistical methods for detecting differentially methylated loci and regions. Front Genet 2014;5:324.
59. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 1997;13:335–40.
60. Woloszynska-Read A, Zhang W, Yu J, Link PA, Mhawech-Fauceglia P, Collamat G, et al. Coordinated cancer germline antigen promoter and global DNA hypomethylation in ovarian cancer: association with the BORIS/CTCF expression ratio and advanced stage. Clin Cancer Res 2011;17:2170–80.
61. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Chan AT, Schernhammer ES, et al. A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. J Natl Cancer Inst 2008;100:1734–8.
62. Bidard F-C, Kiavue N, Ychou M, Cabel L, Stern M-H, Madic J, et al. Circulating tumor cells and circulating tumor DNA detection in potentially resectable metastatic colorectal cancer: a prospective ancillary study to the unicancer prodige-14 trial. Cells 2019;8:516.
63. Cabel L, Decraene C, Bieche I, Pierga J-Y, Bennamoun M, Fuks D, et al. Limited sensitivity of circulating tumor DNA detection by droplet digital PCR in non-metastatic operable gastric cancer patients. Cancers (Basel) 2019;11:396–10.
64. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA methylation clocks in aging: categories, causes, and consequences. Mol Cell 2018;71:882–95.
65. Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet 2007;23:413–8.
66. Pisanic TR, Asaka S, Lin S-F, Yen T-T, Sun H, Bahadirli-Talbott A, et al. Long interspersed nuclear element 1 retrotransposons become deregulated during the development of ovarian cancer precursor lesions. Am J Pathol 2019;189:513–20.
67. Taylor MS, Connie W, Fridy PC, Zhang SJ, Senussi Y, Wolters JC, et al. Ultrasensitive detection of circulating LINE-1 ORF1p as a specific multi-cancer biomarker. Cancer Discov 2023;13:2532–47.
68. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun 2017;8:324.
69. Douville C, Cohen JD, Ptak J, Popoli M, Schaefer J, Silliman N, et al. Assessing aneuploidy with repetitive element sequencing. Proc Natl Acad Sci U S A 2020;47:201910041.
70. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899–905.
71. Kinde I, Papadopoulos N, Kinzler KW, Vogelstein B. FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing. PLoS One 2012;7:e41162.
72. Belic J, Koch M, Ulz P, Auer M, Gerhalter T, Mohan S, et al. Rapid indentification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. Clin Chem 2015;61:838–49.
73. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, et al. Large-scale structure of genomic methylation patterns. Genome Res 2006;16:157–63.
74. Umetani N, Giuliano AE, Hiramatsu SH, Amersi F, Nakagawa T, Martino S, et al. Prediction of breast tumor progression by integrity of free circulating DNA in serum. J Clin Oncol 2006;24:4270–6.
75. Lennon AM, Buchanan AH, Kinde I, Warren A, Honushefsky A, Cohain AT, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. Sci 2020;369:eabb9601.
76. Douville C, Lahouel K, Kuo A, Grant H, Avigdor BE, Curtis SD, et al. Machine learning to detect the SINEs of cancer. Sci Transl Med 2024;16:eadi3883.
77. Annapragada AV, Niknafs N, White JR, Bruhm DC, Cherry C, Medina JE, et al. Genome-wide repeat landscapes in cancer and cell-free DNA. Sci Transl Med 2024;16:eadj9283.
78. Zhou Q, Kang G, Jiang P, Qiao R, Lam WKJ, Yu SCY, et al. Epigenetic analysis of cell-free DNA by fragmentomic profiling. Proc Natl Acad Sci U S A 2022;119:e2209852119.