Title: Non-invasive multi-cancer diagnosis using DNA hypomethylation of LINE-1 retrotransposons

Authors: 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], Caroline Hego[4], Aurore Rampanou[4], Constance Lamy[6], Maud Kamal[6], Christophe Le Tourneau[6], Mathieu Séné[7], Ivan Bièche[7], Cecile 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], Charlotte Proudhon[1,4,5,*].

Affiliations:

[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.

† Equal contribution

* Corresponding author: charlotte.proudhon@inserm.fr

Abstract:

The detection of circulating tumor DNA, which allows non-invasive 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. We implemented a new highly sensitive strategy to detect base-pair resolution methylation patterns from plasma DNA and assessed the potential of hypomethylation of LINE-1 retrotransposons as a non-invasive multi-cancer detection biomarker. Resulting machine learning-based classifiers showed powerful correct classification rates discriminating healthy and tumor plasmas from 6 types of cancers in two independent cohorts (AUC = 88% to 100%, N = 747). This should lead to the development of more efficient non-invasive diagnostic tests adapted to all cancer patients, based on the universality of these factors.

One-Sentence Summary: LINE-1 retrotransposons hypomethylation is a sensitive and specific biomarker to detect multiple forms of cancer non-invasively.

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 non-invasively [4,5] and demonstrated the potential of circulating tumor DNA (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 cancer patients 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 towards 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 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 cancer patients [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, since they largely focus on cancer hypermethylation and unique sequences, it involves targeting specific regions for each cancer subtype. 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]. To obtain a global representation of the hypomethylation occurring during carcinogenesis and to increase sensitivity, we chose to target retrotransposons of the Long-Interspersed Element-1 family (L1) and in particular primate-specific copies (L1PA).

These elements have tens of thousands of copies per cell and are hypomethylated in multiple cancers [42]. Two studies have explored L1 global methylation profiles from plasma [43,44] 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 base-pair 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 DIAMOND (for **D**etection of Long **I**nterspersed Nuclear Element **A**ltered **M**ethylation **ON** plasma **D**NA). 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.

Results

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

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 base-pair resolution analysis and designed a multiplexed PCR based on 8 amplicons covering L1PAs (**Fig. 1A, Table S1**, **Fig. S1A**). We detected thousands of L1PA elements scattered throughout the genome as shown by the genomic hits obtained from a healthy plasma, an ovarian tumor, and a uveal melanoma tumor sequenced at high depth (**Fig. 1B, Table S2**). We observed similar profiles for the three samples, as well as for healthy and cancer plasmas with standard coverage (**Fig. S1B-E**). This demonstrated the robustness of the approach. Overall, the estimated number of L1PA targets is about 30-40,000 elements per genome including half of the human specific copies (L1HS) and many copies of the other L1PA subfamilies (**Fig. 1C, Table S2**). This represents an estimate of 87-120,000 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 (DMR) calling [45]. We, thus, developed a new computational pipeline to accurately align repetitive sequencing data without using a reference genome (**Fig. S1F**). 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 (**Fig. S2A-B**). As expected, the 5' end of the L1 copies targeted is heavily methylated [42,46], particularly within the 2nd amplicon. We also observed quite high levels in both the 5th 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'UTR. 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 OVC, BRC and UVM tumors compared to healthy white blood cells and healthy tissues collected adjacent to ovarian tumors (**Fig. 2A**). Next, we tested a cohort of 473 plasma samples including 123 healthy controls and samples from patients with 6 different types of cancer, covering metastatic (M+) and localized (M0) stages (**Table S3**). This includes colorectal and ovarian cancers in which a substantial rate of L1 hypomethylation has previously been reported [47,48]. 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, 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). Hence, 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 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 PCA analysis (**Fig. S2C**). 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 unique molecular identifiers (UMIs) into the library (**Fig. S1A**). Based on the combination of the 30 CpG targets divided into their 7 amplicons, we extracted a total of 372 unique features (**Fig. S2D**). 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 (**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 PCA analysis shown in **Fig. S2E**, underlining the contribution of highly methylated haplotypes towards the healthy group versus the lowly methylated haplotypes separating cancer samples (middle panel). This separation involves haplotypes from all amplicons (right panel). These results demonstrate that L1 hypomethylation can robustly be observed from cancer plasma DNA at the level of single CpG sites but also at the level of 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, and assessed their performances to automatically separate healthy from tumor plasmas. By testing all cancer samples without subtype specification, the methylation of L1PA elements showed an extremely good ability to discriminate between healthy and tumor plasmas, with an overall area under the curve (AUC) of 94% for both types of features (**Fig. 3A and 3C**). 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 98-100% (**Fig. 3B-C**). 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% with both types of features). 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 cancer to detect non invasively, as reported in previous liquid biopsy multi-cancer tests [11,39,40]. Overall, we observed similar results using single-CpGs methylation levels or using haplotype features. This can be explained by the high correlation observed between these 2 types of features (**Fig. S3A**). Subsequently, we evaluated the importance of the features used by our classifiers (**Fig. 3E-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+). The haplotypes covering these positions showed similar patterns (**Fig. 3F**). Haplotypes provide a more detailed view of the methylation patterns with a strong importance of the most methylated or non-methylated 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, #8 in BRC M0 and LC M+). Overall, this suggests that L1PA methylation alterations 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 3 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 3 stage categories (AUC_{Early} = 95%, $AUC_{Adv.}$ = 97%, AUC_{Meta} = 95%; **Fig. 3G-H, S3B**) with a mean sensitivity of 70% for early stages, (Sen_{Early} = 70%, $Sen_{Adv.}$ = 90%, Sen_{Meta} = 69%; **Fig. 3I**). Strikingly, L1PA methylation 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 59% for ovarian cancer (unpublished data), 38% for colon

cancer [49] and 52% for metastatic breast cancer [13,14] (**Fig. 3J**). We particularly achieved remarkable performance on the cohort of 27 localized gastric cancers with a detection rate of 95% of true positive as compared to 12% for mutation screening [50]. This is mostly due to the fact that methylation changes occur in virtually all cancer patients, unlike recurrent mutations.

Multi-cancer 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, 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**). While methylation at single-CpG within cancer subgroups showed slightly more variability (**Fig. S4A**), global methylation levels were quite reproductible between the two cohorts, showing similar distributions and no statistical differences (**Fig. 4D**), except for non-metastatic ovarian cancers. There was an important heterogeneity among the OVC M0 samples of cohort 2, which clustered into two distinct groups, while cohort 1 was more homogeneous (**Fig. S4B**). Notably, no correlation was found with available clinico-histopathological parameters (age, staging, CA125 level, mutational status, treatment or response to therapy). Differential haplotype proportions between healthy and cancer subgroups where also mostly conserved (**Fig. S4C, Table S8**). Overall, the method showed good reliability with the 7-amplicon panel used and good robustness in detecting L1 methylation levels and changes. Since age-related changes in DNA methylation have been described [51,52] and that the healthy donors included in the study are younger overall than the cancer patients (**Fig. S5A**), we have investigated whether there was an effect on the methylation patterns we studied. We found a significant but very small effect which appeared much smaller that the effect of disease status (**Fig. S5B-C**). This small effect was tending towards an increase in methylation with age (**Fig. S5D**). Furthermore, we observed similar patterns and differences between healthy and cancer samples when adjusting for the age (**Fig. S5E-G**), 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 histological type (**Fig. 4E,G**), and AUC between 88%-100% for the 'cancer-types' models (**Fig. 4F,G**). We observed again great sensitivities at 99% specificity (**Fig. 4H**) with notably 54% for localized breast cancer. It was, however, lower for metastatic lung cancer with a mean sensitivity of 49%. We observed that haplotype models were more robust compared to single-CpG methylation rates (**Fig. S4D-G**). This could be explained by the fact that haplotypes consist of 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 3-stage classifiers and observed great classification performances with a mean AUC of 99% (**Fig. 4I-J**) and a mean sensitivity of 78% for early stages (**Fig. 4K**). 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.

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

We detected significantly more hypomethylation for more advanced stages of the disease, in particular in metastatic stages compared to localized stages (**Fig. 5A**). However, there was no

significant differences between metastatic tumor tissues and primary tissues (**Fig. 5B**), which confirmed that L1PA methylation alteration is an early event in carcinogenesis [54,55] and also affects early-stage cancers. The differences observed in the blood reflect the ctDNA fraction, which is known to correlate with the tumor burden [1,9]. This demonstrates the quantitative potential of DIAMOND, which could help quantify the tumor burden and monitor the disease. A recognized marker to non-invasively estimate the tumor burden and the fraction of ctDNA is the aneuploidy or copy number alterations (CNA) [56], a hallmark of cancer genomes [57]. 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 [58]. 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 copy number alterations, at the level of chromosome arms as previously described ([58,59], see methods) and obtained similar results to those found with CytoScan arrays (**Fig. S6A**). We observed low alteration scores for the normal-like breast cell line HTERT-HME1 and good correlations between the 2 methods for the majority of the cell lines (**Fig. S6B**). 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 subgroups z-scores mirrored global hypomethylation profiles (**Fig. 5E, 2B and 4D**), both reflecting tumor burden and ctDNA fractions available. However, global methylation rates and z-scores were only moderately anti-correlated (**Fig. 5F**), demonstrating that these are partially independent markers that can provide distinct signals (**Fig. S6C**). To obtain a final classification labelling each sample as healthy or cancer, we used a 2-step categorization incorporating CNA analysis, which improved cancer detection. We used the probability of the cancer prediction provided by the methylation-based validation model, 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 methods and **Fig. S6C**). This classifier achieved high sensitivities with specificities between 97-100% for 5 distinct cancer types (BRC, CRC, GAC, LC, OVC, **Fig. 5G**) and could be applied as is in the clinic. This was particularly promising for localized breast cancer with a sensitivity of 100% and a specificity of 100%.

Discussion

In this study, we established a robust proof of concept that targeting hypomethylation of retrotransposons from cell-free DNA is a sensitive and specific biomarker to detect multiple forms of cancer non-invasively. Repetitive regions provide genome-wide information as they hold half of the CpG sites present in the human genome [60]. Hypomethylation of L1 elements, which is a common feature of multiple forms of cancer, help cover the heterogeneous profiles of cancer patients in a single test. Previous studies have left these regions aside as they are inherently difficult to map, and DMR analysis is commonly performed on mapped data. We have developed a new pipeline to detect methylation profiles at repeats with a single base-pair resolution, without resorting to mapping on a reference genome. This allowed us to retain most of our data, which 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 competes with recent cfDNA methylation tests, such as the Galleri and CancerSeek tests. DIAMOND targets about 100,000 CpG sites, ten times less compared to the 1,100,000 CpGs targeted by Galleri [39,40]. Nonetheless, we reached similar or higher levels of sensitivity in 4 of the 5 cancers tested with the 3 methods. Notably, we achieved a 95% sensitivity on localized stages of gastric cancers compared to 47% reached with Galleri and 72% with CancerSeek. We also achieved a 54% sensitivity on localized stages of breast cancers, compared to 28% achieved by Galleri and 33% by CancerSeek. 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. We therefore anticipate that our method has the potential to be applied for the development of routine clinical tests. To push the DIAMOND assay towards a clinically applicable test, we also demonstrated that DIAMOND data can be used to perform copy number alterations 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. Further testing with a larger number of samples covering earlier stages, 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. It will also be important to study the impact of other conditions, such as auto-immune 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[55] demonstrated a high specificity and no sign of L1 reactivation in blood of patients with auto-immune 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.

References

1. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science Translational Medicine* **6**, 224ra24-224ra24 (2014).

2. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine* **20**, 548–554 (2014).

3. Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine* **7**, 302ra133 (2015).

4. Bidard, F.-C., Weigelt, B. & Reis-Filho, J. S. Going with the flow: from circulating tumor cells to DNA. *Science Translational Medicine* **5**, 207ps14-207ps14 (2013).

5. Diaz, L. A. & Bardelli, A. Liquid Biopsies: Genotyping Circulating Tumor DNA. *Journal of Clinical Oncology* **32**, 579–586 (2014).

6. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid biopsy enters the clinic — implementation issues and future challenges. *Nature Reviews Clinical Oncology* 1–16 (2021) doi:10.1038/s41571-020-00457-x.

7. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics* **20**, 1–18 (2019).

8. Stover, D. G. *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.* **36**, 543–553 (2018).

9. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **17**, 223–238 (2017).

10. Mattox, A. K. *et al.* Applications of liquid biopsies for cancer. *Science Translational Medicine* **11**, eaay1984 (2019).

11. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **1**, eaar3247-10 (2018).

12. Riva, F. *et al.* Patient-Specific Circulating Tumor DNA Detection during Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. *Clin Chem* **63**, 691–699 (2017).

13. Jeannot, E. *et al.* A single droplet digital PCR for ESR1 activating mutations detection in plasma. *Oncogene* **73**, 1–9 (2020).

14. Darrigues, L. *et al.* Circulating tumor DNA as a dynamic biomarker of response to palbociclib and fulvestrant in metastatic breast cancer patients. *Breast Cancer Res.* 1–10 (2021) doi:10.1186/s13058-021-01411-0.

15. Silveira, A. B. *et al.* Multimodal liquid biopsy for early monitoring and outcome prediction of chemotherapy in metastatic breast cancer. *npj Breast Cancer* 1–9 (2021) doi:10.1038/s41523-021-00319-4.

16. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. *Nature Reviews Cancer* **11**, 726–734 (2011).

17. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380-10 (2017).

18. Pol, Y. van der & Mouliere, F. Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA. *Cancer Cell* **36**, 350–368 (2019).

19. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics* **8**, 286–298 (2007).

20. Feinberg, A. P. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. *N Engl J Med* **378**, 1323–1334 (2018).

21. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature Genetics* **49**, 635–642 (2017).

22. Li, W. *et al.* CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Research* **46**, e89–e89 (2018).

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* **9**, 26 (2015).

24. Chan, K. C. A. *et al.* Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18761–18768 (2013).

25. Legendre, C. *et al.* Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clinical Epigenetics* **7**, 1–10 (2015).

26. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E5503–E5512 (2015).

27. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proceedings of the National Academy of Sciences* 201519286–34 (2016) doi:10.1073/pnas.1519286113.

28. Garrigou, S. *et al.* A Study of Hypermethylated Circulating Tumor DNA as a Universal Colorectal Cancer Biomarker. *Clinical Chemistry* **62**, 1–11 (2016).

29. Barault, L. *et al.* Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. *Gut* **67**, 1995–2005 (2018).

30. Jin, S. *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.* **118**, e2017421118-8 (2021).

31. Lun, F. M. F. *et al.* Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clinical Chemistry* **59**, 1583–1594 (2013).

32. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).

33. Nassiri, F. *et al.* Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nature Medicine* **26**, 1–12 (2020).

34. Nuzzo, P. V. *et al.* Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nature Medicine* **26**, 1–11 (2020).

35. Liu, X. *et al.* Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by methylated CpG tandem amplification and sequencing (MCTA-Seq). *Clinical Epigenetics* 1–13 (2019) doi:10.1186/s13148-019-0689-y.

36. Luo, H. *et al.* Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Science Translational Medicine* **12**, eaax7533 (2020).

37. Chen, X. *et al.* Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nature Communications* **11**, 1–10 (2020).

38. Cao, F. *et al.* Integrated epigenetic biomarkers in circulating cell-free DNA as a robust classifier for pancreatic cancer. *Clinical Epigenetics* 1–14 (2020) doi:10.1186/s13148-020-00898-2.

39. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).

40. Klein, E. A. *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology* **32**, 1167–1177 (2021).

41. Ross, J. P., Rand, K. N. & Molloy, P. L. Hypomethylation of repeated DNA sequences in cancer. *Epigenomics* **2**, 245–269 (2010).

42. Burns, K. H. Transposable elements in cancer. *Nature Reviews Cancer* **17**, 415–424 (2017).

43. Gainetdinov, I. V. *et al.* Hypomethylation of human-specific family of LINE-1 retrotransposons in circulating DNA of lung cancer patients. *Lung Cancer* **99**, 127–130 (2016).

44. Nagai, Y. *et al.* LINE-1 hypomethylation status of circulating cell-free DNA in plasma as a biomarker for colorectal cancer. *Oncotarget* **8**, 11906–11916 (2017).

45. Robinson, M. D. *et al.* Statistical methods for detecting differentially methylated loci and regions. *Front Genet* **5**, 324 (2014).

46. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).

47. Woloszynska-Read, A. *et al.* Coordinated Cancer Germline Antigen Promoter and Global DNA Hypomethylation in Ovarian Cancer: Association with the BORIS/CTCF Expression Ratio and Advanced Stage. *Clinical Cancer Research* **17**, 2170–2180 (2011).

48. Ogino, S. *et al.* A Cohort Study of Tumoral LINE-1 Hypomethylation and Prognosis in Colon Cancer. *JNCI: Journal of the National Cancer Institute* **100**, 1734–1738 (2008).

49. Bidard, F.-C. *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* **8**, 516–13 (2019).

50. Cabel, L. *et al.* Limited Sensitivity of Circulating Tumor DNA Detection by Droplet Digital PCR in Non-Metastatic Operable Gastric Cancer Patients. *Cancers* **11**, 396–10 (2019).

51. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell* **71**, 882–895 (2018).

52. Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. *Trends Genet.* **23**, 413–418 (2007).

53. Taylor, M. S. & Burns, K. H. Ultrasensitive detection of circulating LINE-1 ORF1p as a specific multi-cancer biomarkers. *bioRxiv* (2023).

54. Pisanic, T. R. *et al.* Long Interspersed Nuclear Element 1 Retrotransposons Become Deregulated during the Development of Ovarian Cancer Precursor Lesions. *Am J Pathology* **189**, 513–520 (2019).

55. Taylor, M. S. *et al.* Ultrasensitive detection of circulating LINE-1 ORF1p as a specific multi-cancer biomarker. *Cancer Discov.* (2023) doi:10.1158/2159-8290.cd-23-0313.

56. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications* **8**, 1–12 (2017).

57. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

58. Belic, J. *et al.* mFast-SeqS as a Monitoring and Pre-screening Tool for Tumor-Specific Aneuploidy in Plasma DNA. in *Circulating Nucleic Acids in Serum and Plasma – CNAPS IX* vol. 924 147–155 (Circulating Nucleic Acids in Serum and Plasma – CNAPS IX, 2016).

59. Belic, J. *et al.* Rapid Identification of Plasma DNA Samples with Increased ctDNA Levels by a Modified FAST-SeqS Approach. *Clin. Chem.* **61**, 838–849 (2015).

60. Rollins, R. A. *et al.* Large-scale structure of genomic methylation patterns. *Genome Research* **16**, 157–163 (2006).

Acknowledgments:

We thank the members of the C.P.'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 NGS platform of the Institut Curie, especially S. Lameiras, V. Raynal and S. Baulande for advice and the non-profit organization "La Vannetaise" for financial support.

Funding:

The NGS 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.P. 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 (CNRS),
The SiRIC 2 Curie program (INCa-DGOS-Inserm_12554),
The DEEP Strive funding (LABEX DEEP 11-LBX0044).
CAA 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).

Author Contributions

MM, MH and CP designed the study
MM, MH, AM, CH, MG, VD, MS, CR and DG performed the experiments
MM, MH, AM, KDS, KVG, CAA and CP analyzed the data.
MM, AM, KDS, MG and KVG performed the statistical analyses.
CH, AR, FCB, JYP, CL, MK, CLT, IB, MHS, OL, LC contributed with identification of clinical samples.
MM, MH, AM, KDS, MG, CAA and CP wrote the manuscript.
All authors participated in revising the manuscript and approved this final version.

Competing Interests statement

CP, MM, MH, and CAA have an ongoing patent application relating to circulating tumor DNA analysis.

Figure 1. Michel *et al*

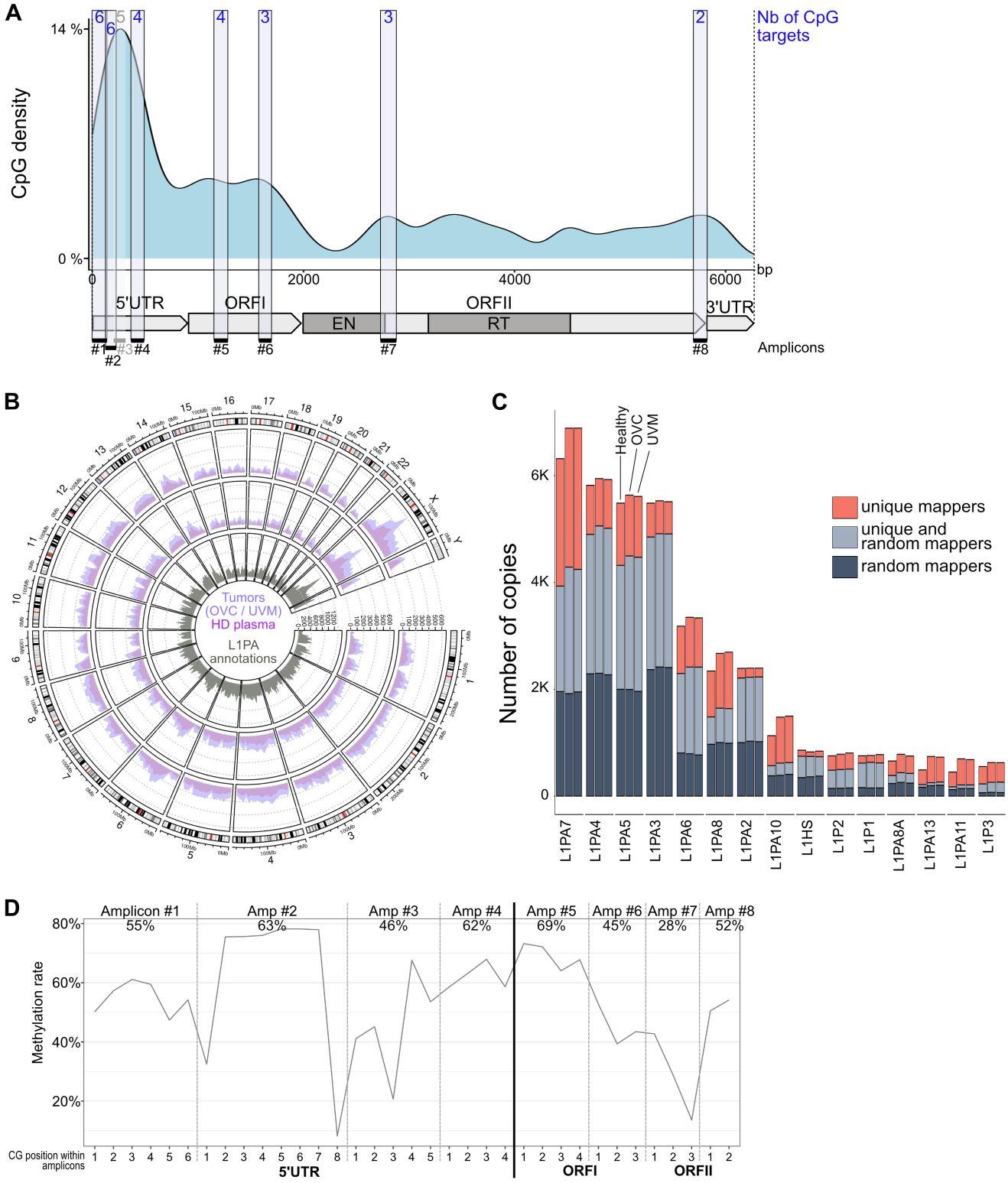


Fig 1. Targeting primate-specific LINE-1 elements reveals plasma DNA-methylation patterns genome-wide

A. CpG density along the structure of a human specific LINE-1 (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 versus 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, grey bottom track). **C.** Histogram summarizing the most represented sub-families of L1 targeted by the DIAMOND assay in the 3 *deep sequenced* samples, in descending order (sum of copies across the 3 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 8 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 grey dotted lines. The dark line marks the end of the 5'UTR. Average levels per amplicon are indicated.

Figure 2. Michel *et al*

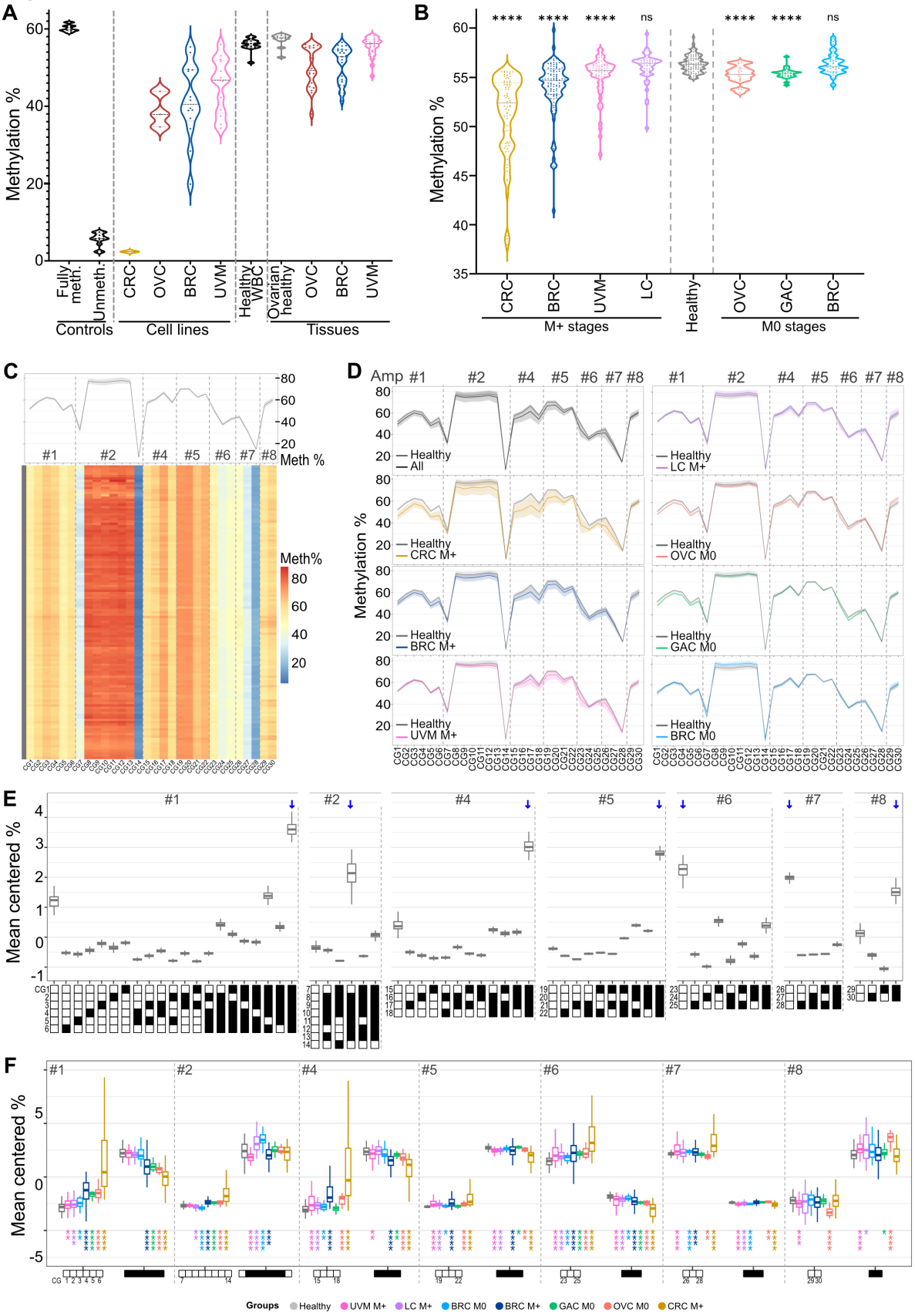


Fig 2. L1PA hypomethylation is detectable from plasma DNA in multiple forms of cancer **A.** Global DNA methylation of fully methylated (SssI-treated DNA, n=13) and unmethylated (Whole-genome amplified DNA, n=12) controls, cancer cell lines or tissues. Ovarian healthy tissues were collected next to ovarian 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. **B.** Global DNA methylation in cancer plasma including metastatic stages (M+) and non-metastatic stages (M0) as well as healthy donor plasmas. Statistical differences between each cancer subgroup and healthy samples were computed using 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, p_{BRC_M0} = 9.10e-01, **Table S4**). Black dotted lines represent the median. **C.** Methylation level at each targeted CpG sites (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 **Fig3F** 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 to healthy donors. Statistical significances were computed using Mann–Whitney *U* test on raw haplotype proportions (**Table S5**).

Figure 3. Michel *et al*

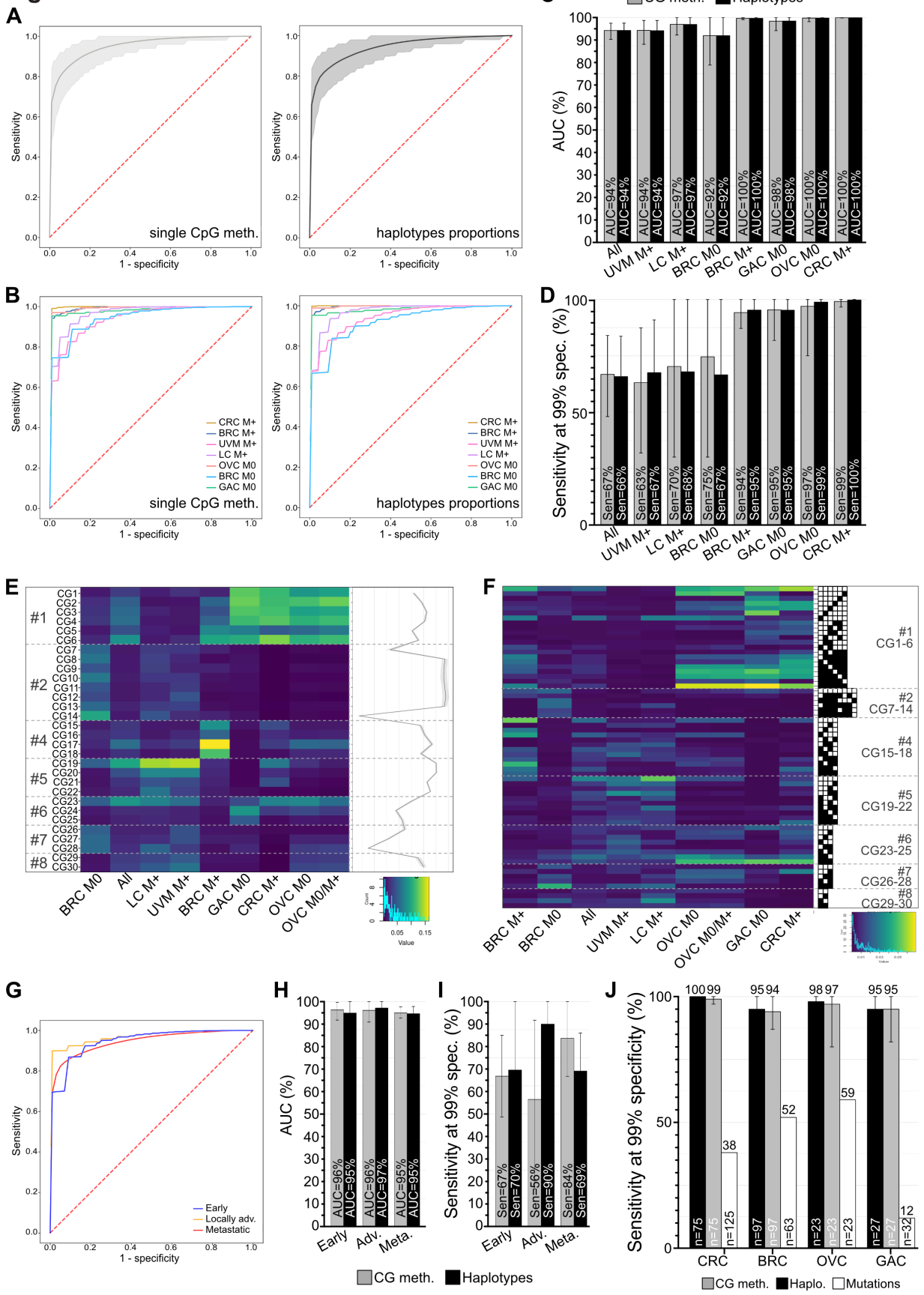


Fig 3. L1PA hypomethylation-based classifiers recognize samples from multiple forms of cancers

A-B. Receiver Operating Characteristic (ROC) curves obtained for plasma samples 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 5000 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 **FigS3C-D**). 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, N_{BRC_M0} = 23 tested versus 123 healthy donors. ROC curves shown are obtained by averaging the sensitivity and specificity of each repetition of learning. **C-D.** Performances for classifiers using single CpG methylation levels (grey) or haplotype proportions (black) presented as AUCs (**C**) or sensitivities at 99% specificity (**D**). Average AUCs are computed from the 5000 AUCs generated by each repetition of learning. Bars indicate 95% CI. **E-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 samples classification with the 3-stage model, using haplotype features. **H-I.** Performances for the 3-stage classifiers using single CpG methylation levels (grey) 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]).

Figure 4. Michel *et al*

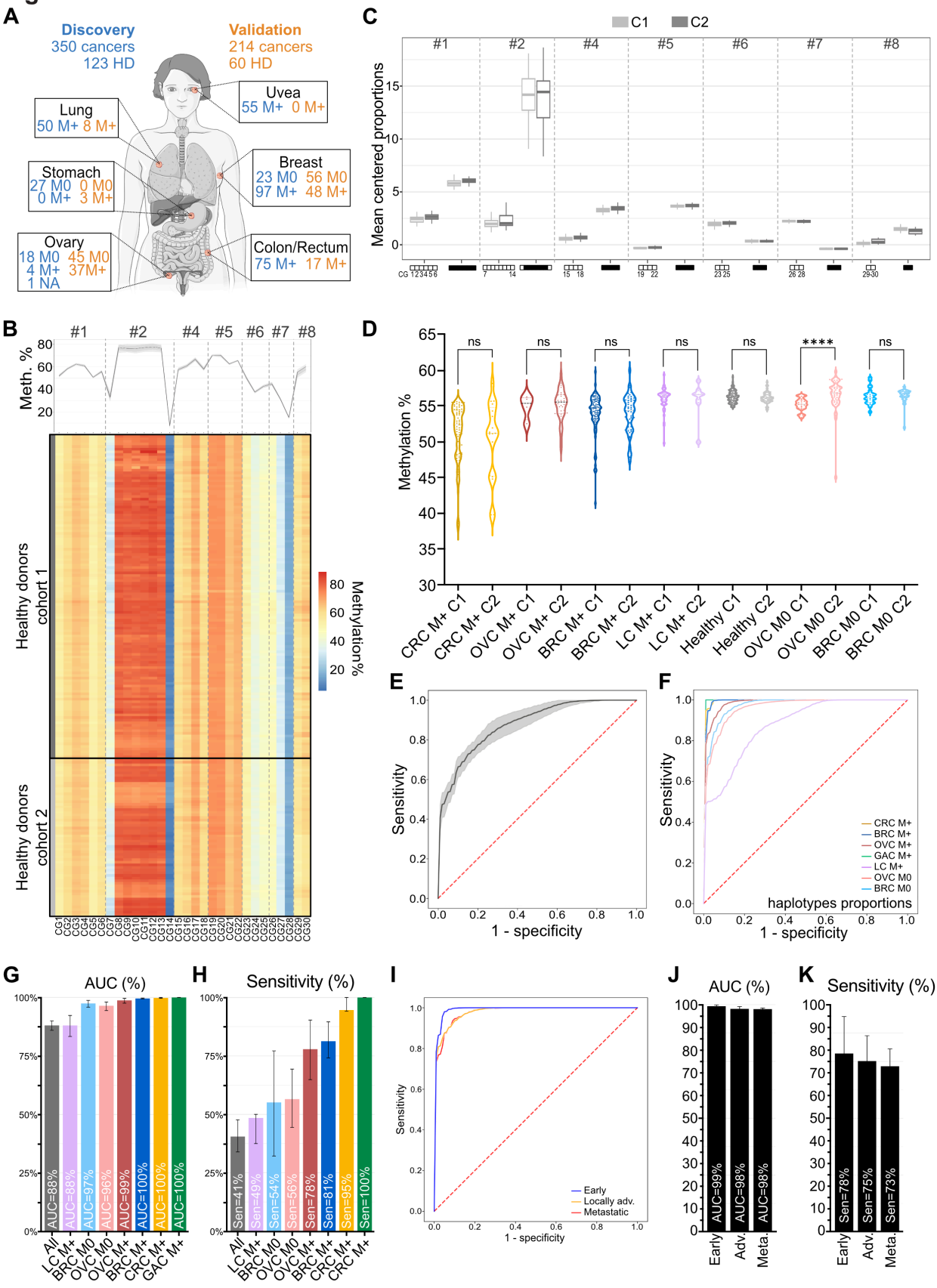


Fig 4. Multi-cancer classification performances are reproducible on an independent cohort **A.** Number of patients and healthy donors (HD) in the discovery cohort (C1) and in the validation cohort (C2) for each cancer type and dissemination stage (non-metastatic: M0 vs metastatic: M+, NA: stage not 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 comes ordered by targeted CpG site on the x-axis (amplicon numbers are indicated). The metaplots represent the average levels for donors of C1 versus C2 at each CpG site. **C.** Mean centered abundance of the most important haplotypes, with the highest co-methylation patterns, in healthy donors from C1 vs C2 (Statistical differences computed using Mann–Whitney *U* test are available in **Table S6**). **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 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, **Table S7**). Black doted lines represent the median. **E-F.** ROC curves obtained for plasma samples classification in the validation cohort with the 'all cancers' model (**E**) or the 'cancer-types' models (**F**) using haplotypes features. All classifications include 5000 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. **G-H**. Performances for validation classifiers using haplotype features presented as AUCs (**G**) or sensitivity at 99% specificity (**H)**. Average AUCs are computed from the 5000 AUCs generated by each repetition of learning. Bars indicate 95% CI. **I-K.** Performances for 3-stage 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 (**I**), AUCs (**J**), or sensitivity at 99% specificity (**K**).

Figure 5. Michel *et al*

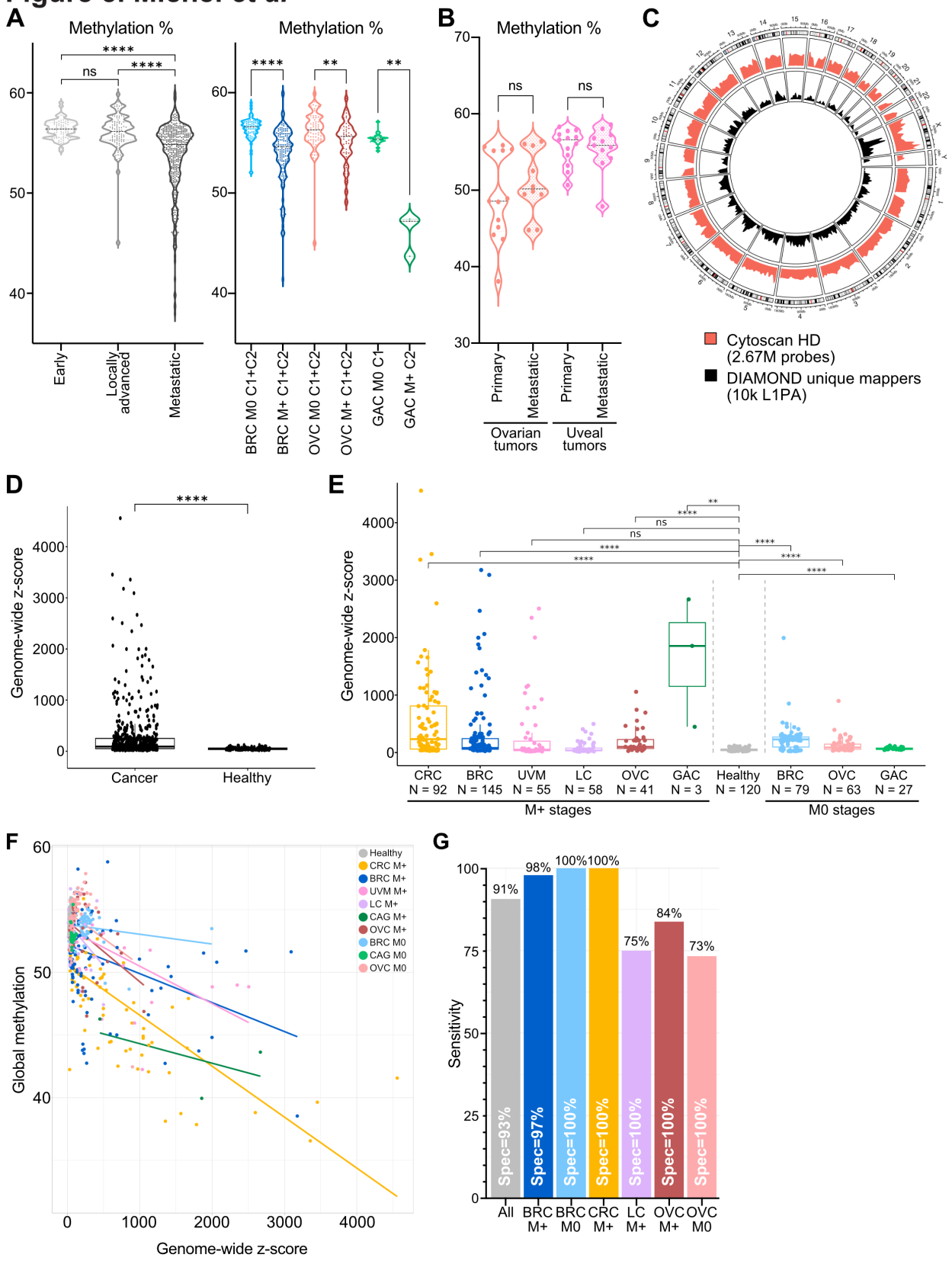


Fig 5. DIAMOND data contain signal to infer the tumor burden, which improves cancer detection

A-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, $p_{GAC_M0/M+}$ = 0.005, **Table S10**) or in primary vs metastatic tissues (**B**, p_{OVC} = 0.257, p_{UVM} = 0.820, **Table S11**). **C**. L1PA unique hits obtained for 15 breast cancer cell lines compared to 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 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, p_{GAC_M0} = 8.40e-06, **Table S12**) **F.** Correlation analysis for genome-wide z-score versus global methylation ($r_{overall}$ = -0.62, p = 1.25e-69). **G.** Performances of the 2-step model incorporating CNA with DNA methylation analysis (Classification is done as follow: $Proba_{Cancer}$ ≤ Threshold C1 AND GZ-score ≤ 121: prediction = Healthy; $Proba_{Cancer}$ > Threshold C1 OR GZ-score > 121: prediction = Cancer, see methods).

Methods

Materials

Cell lines

Cell lines screened in Fig. 2A are the following: CRC (HCT116); OVC (SKOV, Caov3, ES-2); BRC (MDA-MB453, SKBR3, MDA-MB361, HCC202, ZR75.1, HCC70, BT474, MDA-MB231, Cal51, MDA-MB157, BT20, MCF7, HCC1954, HCC1569, HCC38); UVM (MP38, MP41, MP46, MP65, MM28, Mel285, Mel270, 92.1, Mel202, omm2.5, Mel290, mm66, omm1).

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 (Paris, France) 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 ovarian cancer or triple-negative breast cancer patients 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), non-metastatic 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 h, to ensure a good quality of cfDNA, by centrifugation at 820 g for 10 min, followed by a second centrifugation of the supernatant at 16,000 g for 10 min and stored at −80°C until use.

Methods

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 (FFPE) 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 µl, 75 µl or 36 µl, respectively.
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 base-pair 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, CA, USA), 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, CA, USA), according to the manufacturer's instructions. Bisulfite-treated DNA was stored at -80°C and further used to build a sequencing library.

Primer design

Eight primer pairs were designed using the LINE-1 Human Specific (L1HS) consensus sequence from Repbase (**Fig. 1A**). Although 5'UTR (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 or PyroMark Softwares. Targeted regions contained 2-7 CpG targets and ranged from 101bp to 150bp, to better capture cfDNA fragments, which have a mean size of 167bp [61], (**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 CS (common sequence) 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 thousands of copies per genome, a high number of distinct UMIs is 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 diversity of sequencing libraries and improve sequencing quality. All primers were obtained from Eurogentec (RP-cartridge purification method).

Preparation of targeted bisulfite sequencing libraries

Sequencing libraries were prepared using three PCR steps (**Fig. S1A**): 1) target-specific linear amplification for UMI assignment, 2) target-specific exponential amplification and 3) barcoding PCR for sample identification. Each library was prepared in two individual reactions (due to the overlap of amplicon 2 with other primers), including: *I)* Multiplex PCR amplification of 7 probes (Amplicon 1, 3, 4, 5, 6, 7, 8), and *II)* Single PCR amplification of amplicon 2.
UMI assignment for multiplex reaction was performed using Platinum™ Multiplex PCR kit Master Mix (Thermofisher, Life Technologies SAS) in a 25 µL reaction containing 1x Platinum™ Multiplex PCR Master Mix, 0.01-0.06 µM mix of reverse primers and up to 5 ng bisulfite-converted DNA at the following thermocycling conditions: 95°C for 5 min followed by 1 cycle at 95°C for 30 s, 58°C for 90 s, 72°C for 40 s. UMI assignment for single reaction was performed using Hot Star Taq Plus DNA Polymerase (Qiagen) in a 25 µL reaction containing 1x Taq PCR Buffer, 0.65 U Hot Star Taq (5U/µL), 0.2 µM dNTPs, 1.5 mM MgCl2, 0.1 µM amplicon 2 reverse primer, up to 4 ng of bisulfite-converted DNA at the following thermocycling

conditions: 95°C for 10 min followed by 1 cycle at 94°C for 60 s, 58°C for 30 s, 72°C for 40 s. To ensure complete removal of the reverse primers and dNTPs, each 25 µL reaction was treated with 50U of Exonuclease I and 10U of FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific) at 37°C for 1 h and heat-inactivated at 80°C for 15 min.
Target-specific exponential amplification for multiple reaction was performed using Platinum™ Multiplex PCR kit Master Mix in a 50 µL reaction containing 1x Platinum™ Multiplex PCR Master Mix, 0.01-0.06 µM mix of forward primers, 0.2 µM CS2 reverse primer and 20 µL of purified PCR product at the following thermocycling conditions: 95°C for 5 min followed by 28 cycles at 95°C for 30 s, 58°C for 90 s, 72°C for 30 s followed by a 10 min 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 1x Taq PCR Buffer, 0.65 U Hot Star Taq (5U/µL), 0.2 µM dNTPs, 1.5 mM MgCl2, 0.2 µM amplicon 2 forward primer, 0.2 µM CS2 reverse primer and 8 ul of purified PCR product at the following thermocycling conditions: 95°C for 10 min, 25 cycles at 94°C for 60 s, 58°C for 30 s, 72°C for 30 s and 10 min 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 1.2x ratio according to the manufacturer's protocol. Purified DNA was eluted in 30 ul of water. Barcoding PCR was performed using universal fluidigm primers. 25 µL of purified pooled PCR product, 1x Phusion HF Buffer, 1 U Phusion Hot Start II DNA Polymerase (Thermo Fisher Scientific), 0.2 µM fluidigm primer, and 0.2 mM dNTPs were mixed in the final volume of 50 µL and amplified with the following conditions: 98 °C for 2 min, followed by 20-25 cycles of 98 °C for 10 s, 62 °C for 30 s, and 72 °C for 30 s followed by a 5 min incubation at 72°C. The amplified product was purified by two consecutive AMPure XP steps using *1)* a low concentration of AMPure XP beads (0.6x – 0.7x ratio) where the beads containing the larger fragments are discarded and supernatant collected (reverse purification) and *2)* higher beads concentration (1.1x – 1.2x ratio) where 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), and pooled equimolarly for sequencing. Sequencing was performed on Illumina HiSeq rapid run mode or NovaSeq (PE 30bp, 170bp).

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 (**Fig. S1A**) were first filtered for reads quality (average >Q20 per read) and then demultiplexed (i.e., cut using *atropos* v1.1.31) 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. 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 (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 (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 haplotypes 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 site. 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 [62] from Python package scikit-learn [63], with the following hyperparameters: n_estimators=300, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_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: *1)* it is less prone to overfitting [62]; *2)* 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 haplotypes data representation [64]; *3)* random forests also inherently return measures of variable importance [62], 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), 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. Model classifications were run 5000 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 [65]. 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 for the 5000 runs. In the case of 'multiclass' classification, the ROC curves of each class were generated by taking the class under consideration as the positive class and the union of all others as the negative class. There is, therefore, no particular weight associated with the healthy plasmas class.

Copy number alterations 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, 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 Log2 combining probes by chromosome arms.
DIAMOND CNA: 1) Z-score calculation: prepossessed reads were uniquely mapped on hg38 genome using Bismark (version 0.23.1). As in Belic *et al.* 2015, only the reads with an alignment score > 15 were kept. Resulting reads from all amplicons (excluding #2 and #3) were merged and normalized number of reads per chromosome arm (excluding sexual chromosomes X and Y) per sample were calculated with R. Next, the amplifications/deletions score was computed using the following formula:

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

with i = a given chromosome arm, n = a given sample and *controls* = a set of reference samples (10 PBMC reference samples for the cell lines, 63 healthy donors 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. 2) 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 ($N_{Healthy}$ = 60, N_{Cancer} = 350) using the genome-wide z-score and calculated the threshold that maximize the sensitivity at 100% specificity.

2-step classification for sample labelling

First, we selected the threshold for the probability of the cancer prediction ($Proba_{Cancer}$) on the discovery cohort maximizing the sensitivity for a 99% specificity, per 'cancer-type' model. We applied this threshold on the $Proba_{Cancer}$ computed with the validation models and reclassified samples which presented a z-score > 121, as cancer ($Proba_{Cancer}$ ≤ Threshold C1 AND GZ-score ≤ 121: prediction = Healthy; $Proba_{Cancer}$ > Threshold C1 OR GZ-score > 121: prediction = Cancer) see **Tables S13-19**.

References

61. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).

62. Breiman, L. Random Forests. *Mach Learn* **45**, 5–32 (2001).

63. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Arxiv* (2012).

64. Boulesteix, A., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov* **2**, 493–507 (2012).

65. Barandela, R., Sánchez, J. S., García, V. & Rangel, E. Strategies for learning in class imbalance problems. *Pattern Recogn* **36**, 849–851 (2003).

Supplementary materials

This includes:

Figs. S1 to S6
Tables S1 to S19 as separated Excel file

Figure S1. Michel *et al*

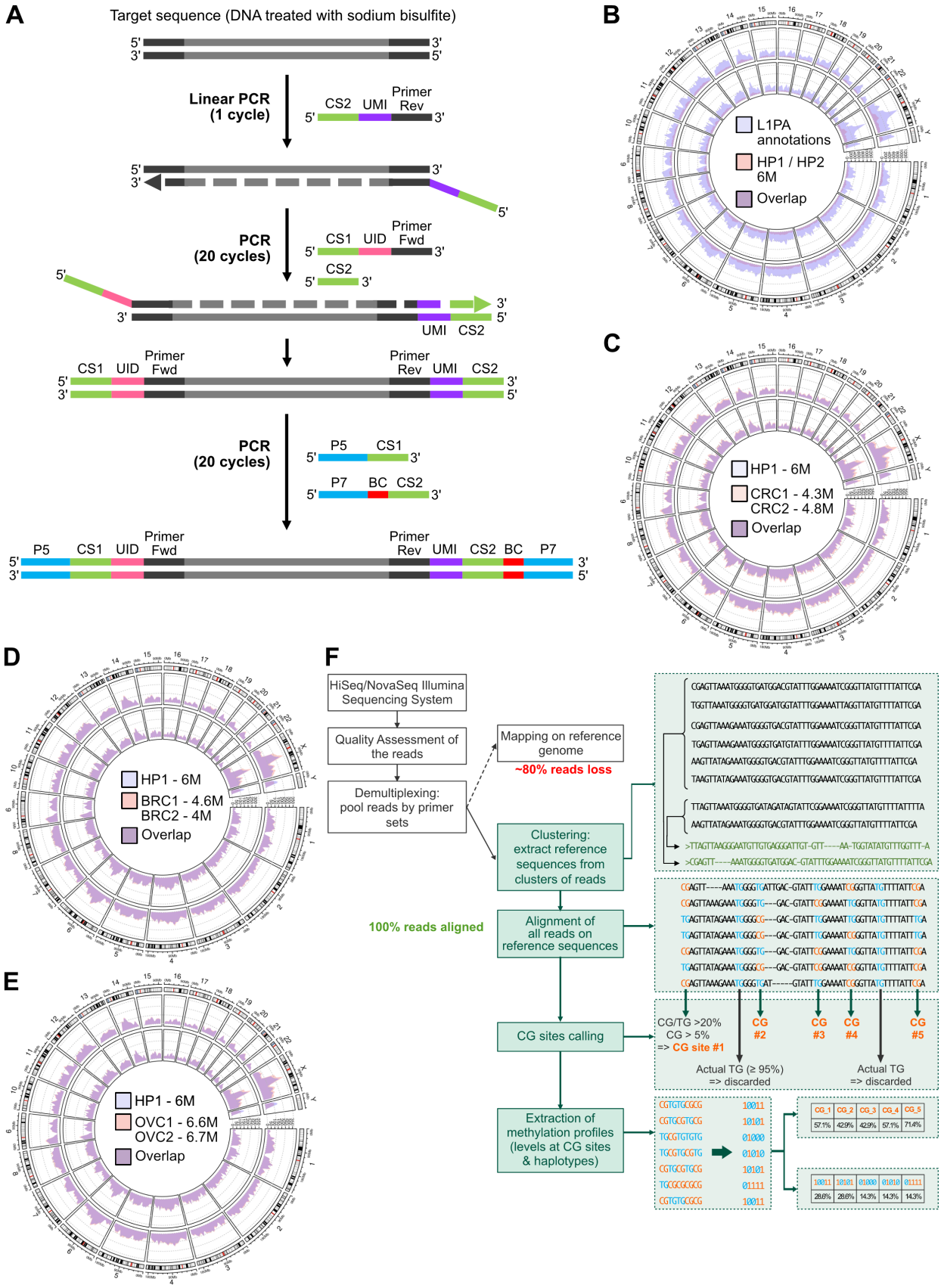
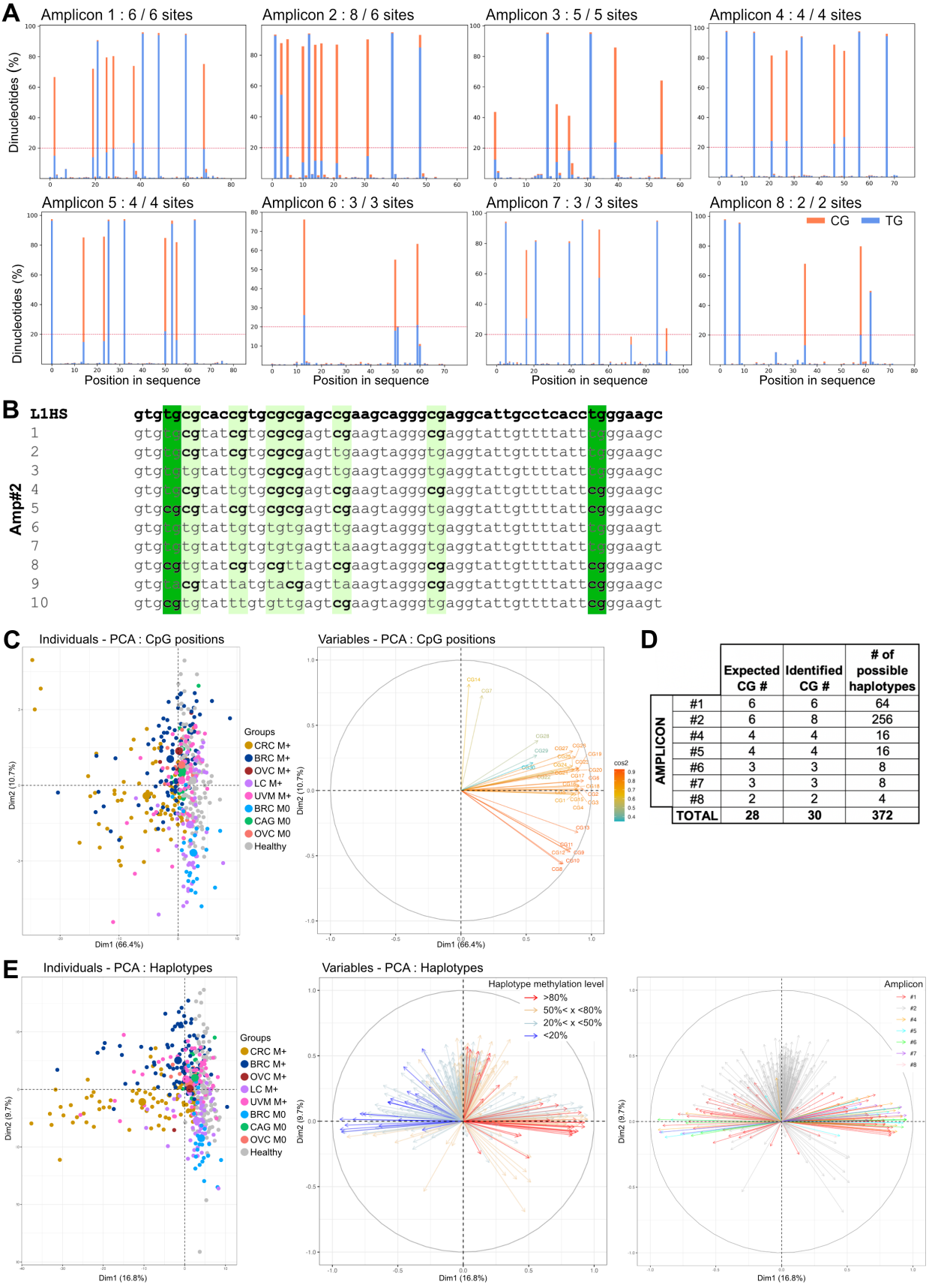


Fig S1

A. Scheme of the targeted bisulfite sequencing strategy used to build the DIAMOND assay libraries. The protocol starts by the incorporation of unique molecular identifiers (UMI) via 1 cycle of linear PCR to identify each initial molecule present in the sample. We also incorporated a 2nd set of molecular identifiers (UID) during the 2nd PCR in order to generate libraries with enough nucleotide diversity which is crucial for a successful downstream sequencing (See method section for more details). **B-E.** Hits obtained across the genome using the DIAMOND assay with standard coverage (indicated in million reads – M) in healthy plasma (HP) over the L1PA annotations **(B)**, in HP versus colorectal cancer (CRC) plasmas **(C)**, in HP versus breast cancer (BRC) plasmas **(D)**, and in HP versus ovarian cancer (OVC) plasmas **(E)**, (see also **Table S2**). **F.** Summary flow chart illustrating the pipeline developed for reference-free alignment of sequencing data (see also Methods).

Figure S2. Michel *et al*



		Expected CG #	Identified CG #	# of possible haplotypes
AMPLICON	#1	6	6	64
	#2	6	8	256
	#4	4	4	16
	#5	4	4	16
	#6	3	3	8
	#7	3	3	8
	#8	2	2	4
	TOTAL	**28**	**30**	**372**

Fig S2

A. Proportion of CG and TG dinucleotides within the 8 regions targeted, after reference-free alignment. Number of CpG sites detected vs expected are indicated for each amplicon. The dashed red line represents the first threshold applied to select dinucleotide positions with >20% (GC+TG). A second threshold of >5% CG is applied. **B.** Alignment of 10 representative sequences of the largest clusters obtained for amplicon 2 relative to the L1HS consensus sequence, highlighting 2 additional CpG positions identified (dark green). **C.** Principal component analysis, based on the average methylation level at each CpG position (n=30), showing the distribution of healthy and cancer samples annotated for their cancer subgroups in the two first dimensions (left panel), and the contribution of CpG positions used as components (right panel). **D.** Number of possible haplotype features extracted from the 30 CpGs within the 7-amplicon panel. **E.** Principal component analysis, based on haplotypes proportions (n=372), showing the distribution of healthy controls and cancer samples annotated for their cancer subgroups in the two first dimensions (left panel), and the contribution of haplotypes used as components (middle and right panels). Middle panel highlights 4 groups of methylation levels relative to the haplotype components. The right panel highlights the contribution of the various amplicons.

Figure S3. Michel *et al*

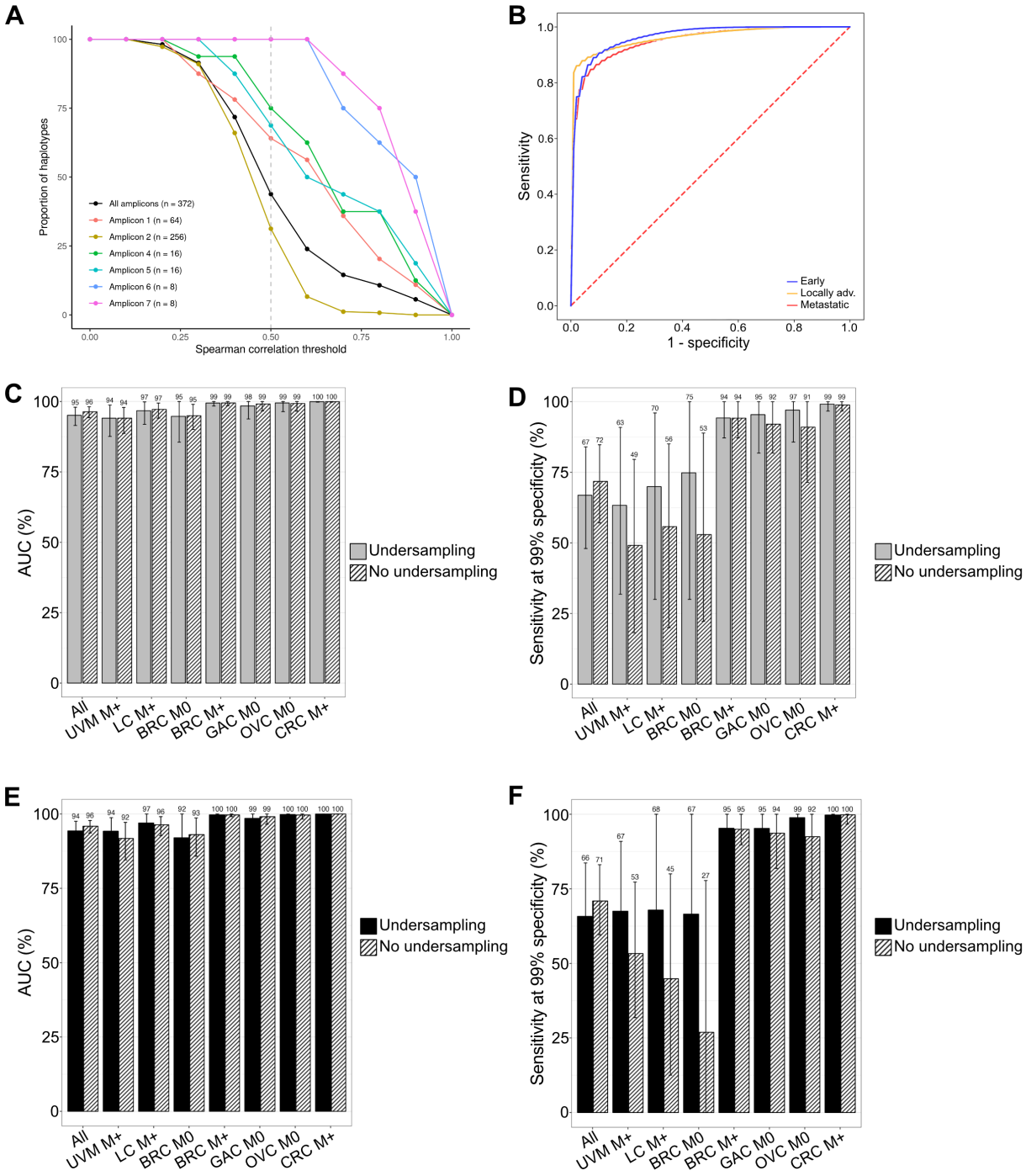


Fig S3.

A. High correlation between haplotype representation and CpG positions shown by the proportion of haplotypes with at least one correlation to CpG sites along spearman rho correlation thresholds. Amplicons 1, 4-8 show high correlation all along. Amplicon 2 shows lower correlations due to its very high number of haplotypes (n = 264). n: number of haplotypes per amplicon. **B.** ROC curves obtained for plasma samples classification with the 3-stage model, using single-CpG methylation rates. **C-F.** Performances for classifiers using single CpG methylation features (**C-D**) or haplotype features (**E-F**) with undersampling (plain bars) or not (hatched bars) presented as AUCs (**C** or **E**) or sensitivities at 99% specificity (**D** or **F**).

Figure S4. Michel *et al*

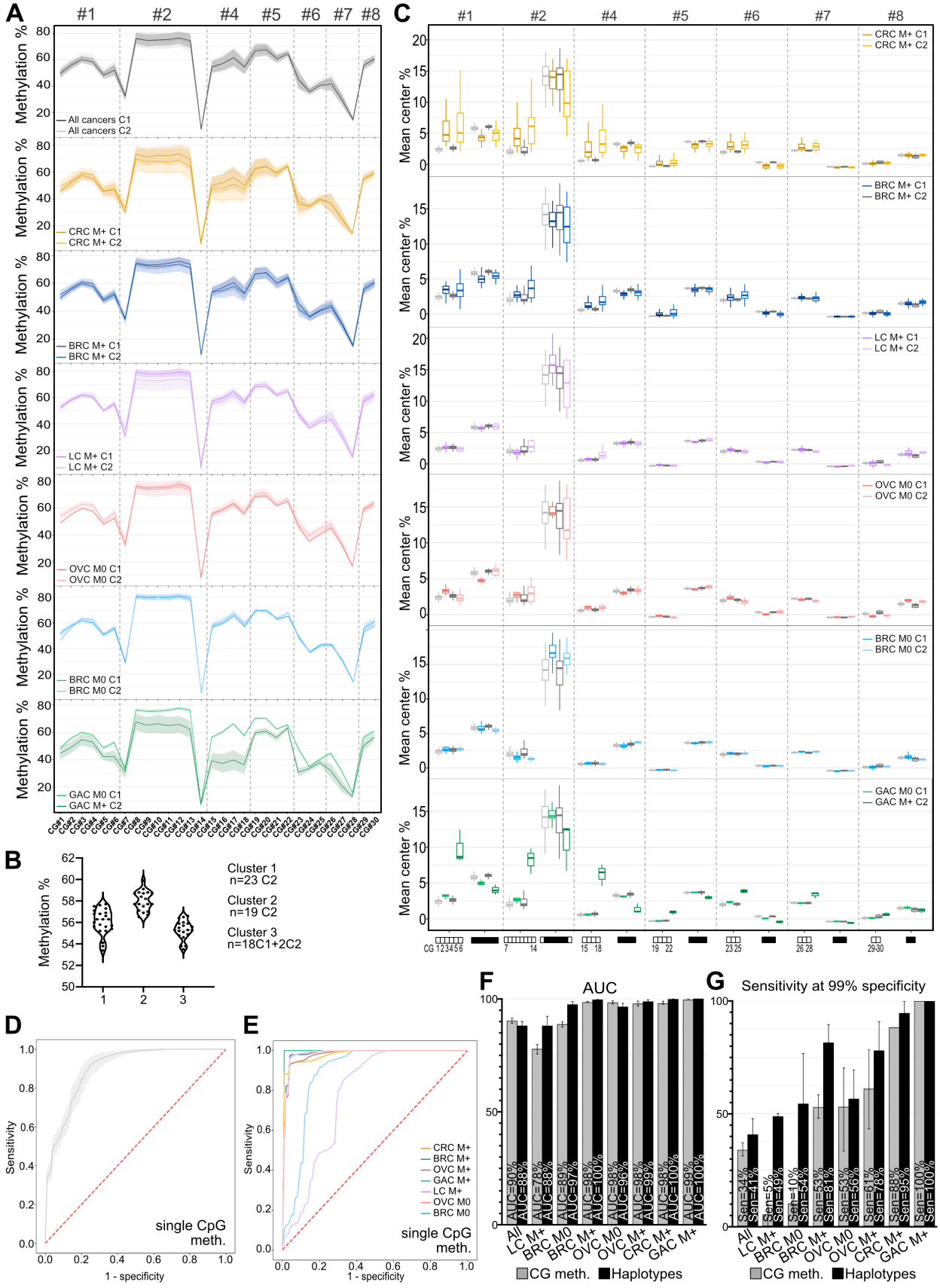
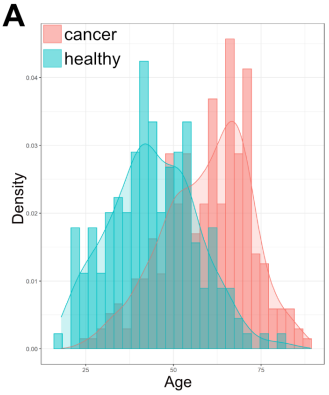


Fig S4.

A. Comparison of the methylation level at individual CpG sites along the L1 for each cancer subgroup in cohort 1 (C1) versus cohort 2 (C2). Amplicon numbers are indicated. **B.** Clustering analysis on OVC M0 stages of cohort 1 (C1) and cohort 2 (C2) showing a split of C2 into cluster 1 and 2 and separation of the whole cohort 1 in a single cluster (cluster 3) including 2 samples of C2. **C.** Comparison of the haplotype proportions for each cancer subgroup and the healthy controls in cohort 1 (C1) versus cohort 2 (C2). Light grey = healthy donors C1, dark grey = healthy donors C2. Statistical analyses are reported in **Table S8**. **D-E.** ROC curves obtained for plasma samples classification in the validation cohort with the 'all cancers' model (**D**) or the 'cancer-types' models (**E**) using single CpG methylation features. All classifications include 5000 stratified random repetitions of learning on the whole discovery cohort and testing on the whole validation cohort without undersampling. **F-G**. Performances for validation classifiers using CpG methylation (grey) or haplotypes (black) features presented as AUCs (**F**) or sensitivity at 99% specificity (**G)**. Average AUCs are computed from the 5000 AUCs generated by each repetition of learning. Bars indicate 95% CI.

Figure S5. Michel *et al*

A

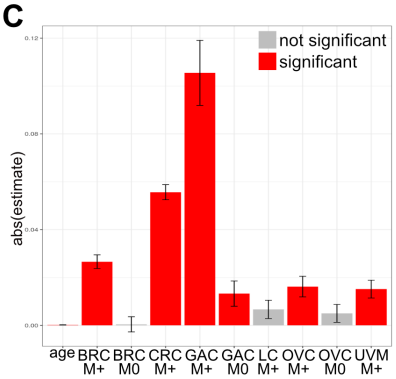


B

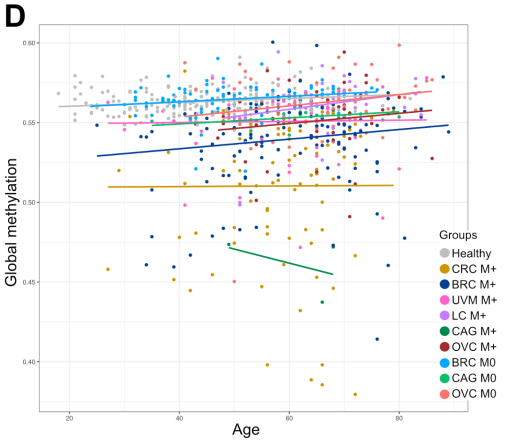
MODEL1	Estimate	Std..Error	t.value	Pr...t..	p.signif
(Intercept)	0.5573	0.0044	125.3515	0.0000	
age	-0.0001	7.7441E-05	-1.7516	0.0803	ns

MODEL2	Estimate	Std..Error	t.value	Pr...t..	p.signif
(Intercept)	0.5549	0.0037	150.1850	0.0000	
age	0.0002	7.4035E-05	2.5678	**0.0104**	*
CRC M+	-0.0556	0.0032	-17.5279	**1.2592E-57**	****
BRC M+	-0.0266	0.0029	-9.3296	**1.2616E-19**	****
UVM M+	-0.0152	0.0037	-4.0715	**5.1840E-05**	****
NSCLC M+	-0.0067	0.0038	-1.7445	0.0815	ns
GAC M+	-0.1055	0.0136	-7.7563	**2.9639E-14**	****
OVC M+	-0.0162	0.0043	-3.7679	**0.0002**	****
OVC M0	-0.0050	0.0037	-1.3506	0.1772	ns
GAC M0	-0.0133	0.0053	-2.5182	**0.0120**	**
BRC M0	0.0005	0.0032	0.1438	0.8857	ns

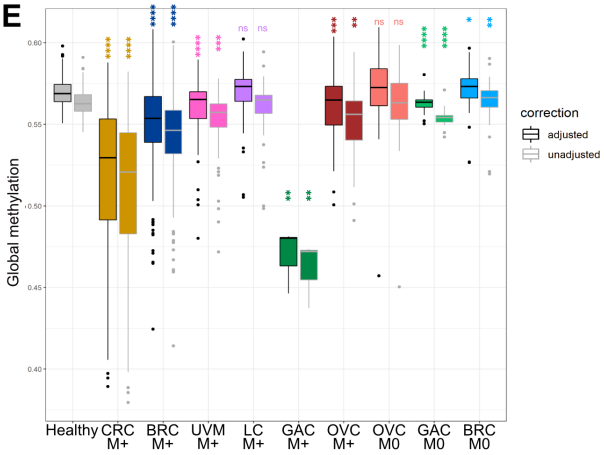
C



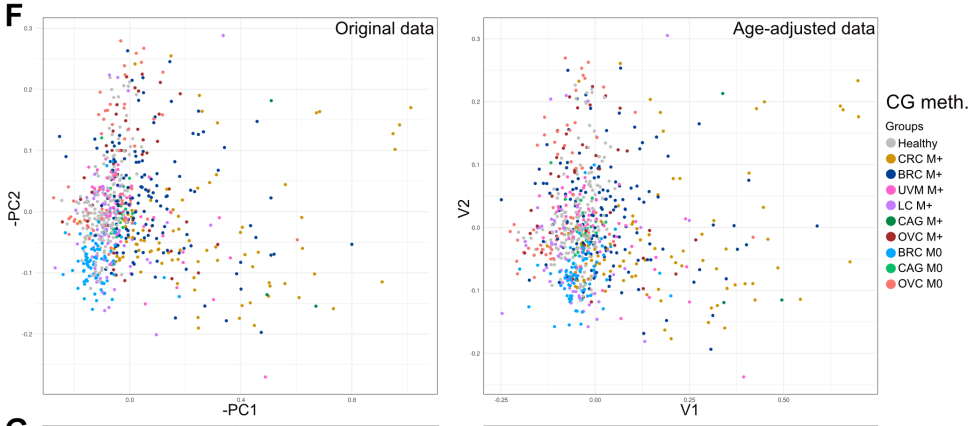
D



E



F



G

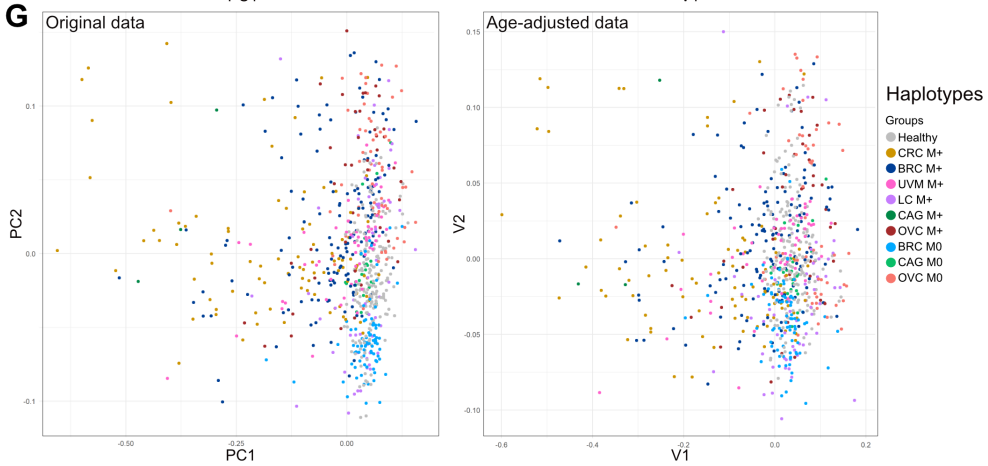
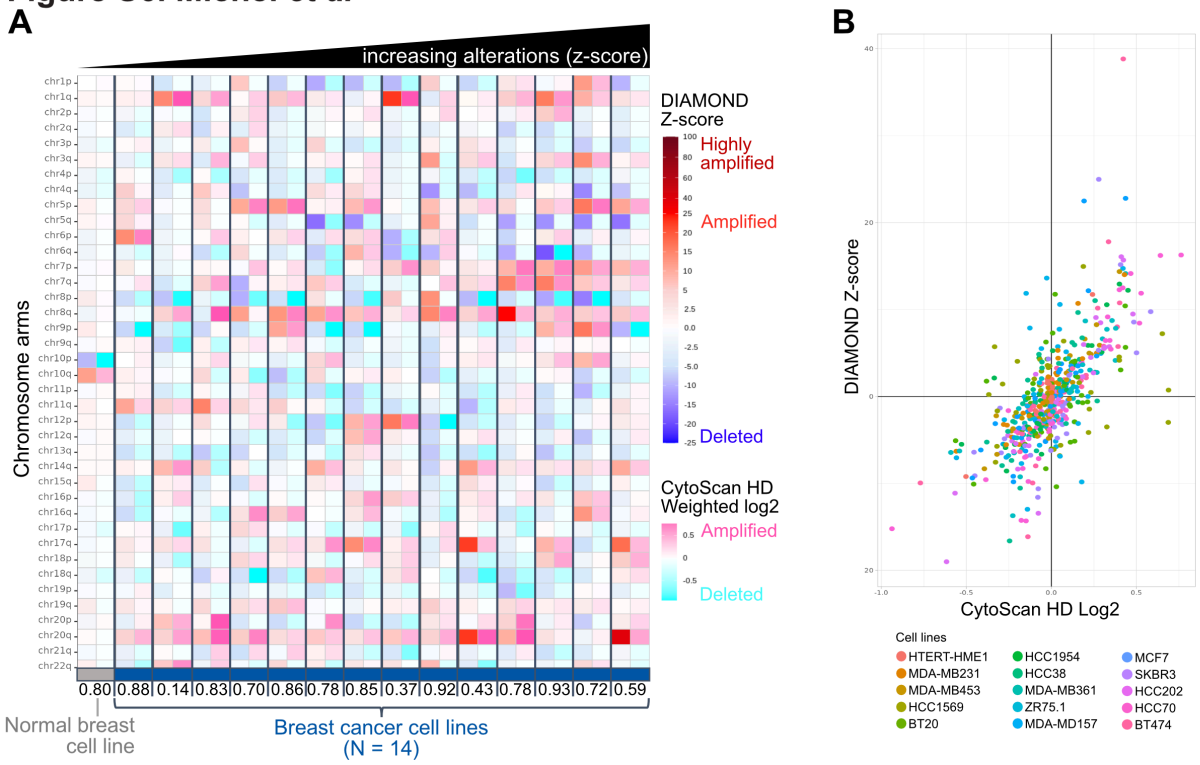


Fig S5.

A. Age distribution of all cancer vs healthy samples (C1 + C2). **B.** Linear models demonstrating that the age did not predict the methylation patterns alone (Model 1) and had a significant but small effect (see also (**C**)) in combination with the 'biological class' representing the cancer subgroups (Model 2). **C.** Impact extent of the parameters tested represented by the absolute values of their Estimate (red: significant, grey: not significant). **D.** Global methylation levels for all samples versus their age. Correlation curves illustrate small impact of age on the methylation patterns in our study, with an increase in methylation with age. **E.** Global methylation levels of original data next to age-adjusted data, organized by subgroups, demonstrating that cancer subgroups versus healthy differences remained similar when adjusting for the age. Statistical differences between each cancer subgroup and healthy samples were computed using Mann–Whitney *U* test (see **Table S9** for detailed p-values and CI). **F-G.** PCA analysis on original data (left panel) or age-adjusted using AC-PCA to adjust for confounding factors conjointly (right panel) for single CpG methylation levels (**F**) or haplotypes (**G**). Adjusting for the age seemed to have limited impact on the data indicating that age is not a confounding factor here.

Figure S6. Michel *et al*



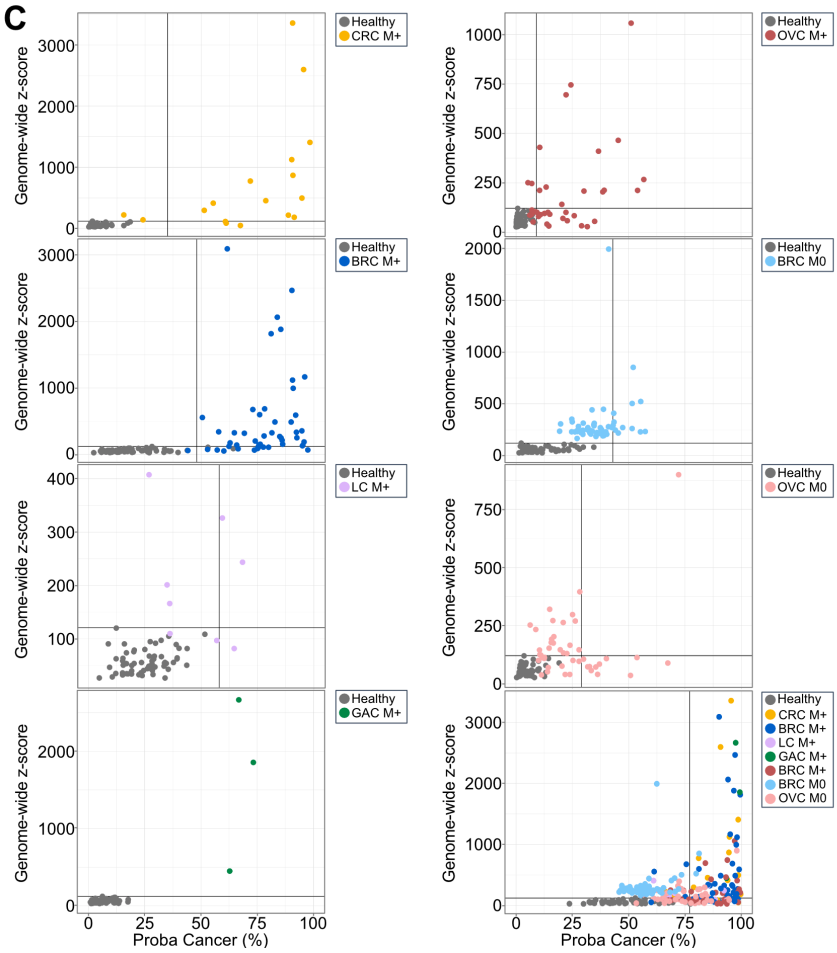


Fig S6.
A. Heatmap comparing copy number alterations (CNA) profiled by DIAMOND Z-score versus Cytoscan HD microarrays for 1 normal-like breast cell line (HTERT-HME1) and 14 breast cancer cell lines (in order: MDA-MB231, MDA-MB453, HCC1569, BT20, HCC1954, HCC38, MDA-MB361, ZR 75.1, MDA-MB157, MCF7, SKBR3, HCC202, HCC70, BT474) organized by increasing aneuploidy measured by the chromosome arms z-scores. Correlation scores are indicated below the heatmap. **B.** Correlation between CNA measured by CytoScan HD microarrays and DIAMOND, $r_{overall} = 0.68$, $p = 8.73e\text{-}80$. **C**. CNA as a function of the probability of a sample to be classified as cancer (Proba Cancer) in the validation models. Graphs for cancer subgroups and all cancers together are shown.

Tables S1 to S19. (Separated excel format file)
Table S1. Targeted bisulfite sequencing primers
Table S2. Target copies and CpG sites relative to the number of sequencing reads
Table S3. Samples list. Sample IDs were not known to anyone outside the research group.
Table S4. Statistical results of global methylation differences between healthy samples and cancer subgroups using Mann–Whitney U test
Table S5. Statistical results of proportion differences between healthy samples and cancer subgroups for the 372 haplotypes using Mann–Whitney U test
Table S6. Statistical results of proportion differences between healthy samples from C1 and C2 for 14 haplotypes selected using Mann–Whitney U test
Table S7. Statistical results of global methyaltion differences between healthy samples and cancer subgroups from C1 and C2 using Mann–Whitney U test
Table S8. Statistical results of proportion differences between healthy samples and cancer subgroups from C1 and C2 for 14 haplotypes selected using Mann–Whitney U test
Table S9. Statistical results of global methylation differences between original data and age-adjusted data using Mann–Whitney U test
Table S10. Statistical results of global methylation differences between cancer stages using Mann–Whitney U test
Table S11. Statistical results of global methylation differences between primary and metastatic tissues using Mann–Whitney U test
Table S12. Statistical results of Genome-wide z-score differences between healthy samples and cancer subgroups using Mann–Whitney U test
Tables S13-S19. Cancer prediction and sample labelling with the 2-step classification integrating CNA analysis - healthy vs each type of model