Comprehensive pan-cancer analysis of cfDNA methylation marks in tumors reveals complex epigenetic regulatory circuits and diagnostic biomarkers

Smiths Lueong
smiths-sengkwawohlueong@uk-essen.de

University Hospital Essen  https://orcid.org/0000-0002-2776-6706

Martin Metzenmacher
Department of Medical Oncology, West German Cancer Center, University Hospital Essen, Essen, Germany  https://orcid.org/0000-0001-5395-9010

Gregor Zaun
Department of Medical Oncology, West German Cancer Center, University Hospital Essen, Essen, Germany

Gina Mayer
Department of Gynecology and Obstetrics, University Hospital Essen, Essen, Germany

Erik Hemmer
Bridge Institute of Experimental Tumor Therapy (BIT), West German Cancer Center, University Hospital Essen, Essen, University Duisburg-Essen, 45147 Essen, Germany

Katharina Lückerath
Department of Nuclear Medicine, University Hospital Essen, University of Duisburg-Essen, Hufelandstraße 55, D-45147 Essen, Germany

Kelsey Pomykala
Institute for Artificial Intelligence in Medicine, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Balazs Hegedues
University Medicine Essen - Ruhrlandklinik  https://orcid.org/0000-0002-4341-4153

Peter Horn
Institute for Transfusion Medicine, University Medicine Essen, University Hospital Essen, University Duisburg-Essen, 45147 Essen, Germany

Marija Trajkovic-Arsic
Bridge Institute of Experimental Tumor Therapy (BIT), West German Cancer Center, University Hospital Essen, Essen, University Duisburg-Essen, 45147 Essen, Germany

Tibor Szarvas
Renata Varaljai
Department of Dermatology, University Hospital Essen, West German Cancer Center, University Duisburg-Essen and the German Cancer Consortium (DKTK), Essen, Germany  https://orcid.org/0000-0001-6852-6859

Corinna Keup
Department of Gynecology and Obstetrics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany  https://orcid.org/0000-0003-3525-7511

Ingeborg Tinhofer-Keilholtz
Department of Radiooncology and Radiotherapy, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin

Stephen George
Department of Radiooncology and Radiotherapy, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin

Sabine Kasimir-Bauer
Department of Gynecology and Obstetrics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Samuel Peña-Llopis
University Hospital Essen  https://orcid.org/0000-0003-3847-1353

Comelius Kürten
Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Essen, University Duisburg-Essen, Essen, Germany

Lukas Boosfeld
Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Essen, University Duisburg-Essen, Essen, Germany

Kirsten Bruderek
Experimental and Translational Research, Department of Otorhinolaryngology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Sven Brandau
University Hospital Essen  https://orcid.org/0000-0002-2702-4163

Christopher Darr
Department of Urology, University Hospital Essen, University of Duisburg-Essen, Hufelandstraße 55, D-45147 Essen, Germany

Thomas Hilser
Department of Medical Oncology, West German Cancer Center, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Viktor Grünwald
Hans Neubauer
Department of Obstetrics and Gynecology, Medical Faculty and University Hospital of the Heinrich-Heine University, Düsseldorf, Germany  https://orcid.org/0000-0002-3467-4105

Irene Esposito
Institute of Pathology, University Hospital Duesseldorf, Heinrich-Heine University, Duesseldorf, Germany

Tanja Fehm
Klinik für Geburtshilfe und Frauenheilkunde, Universitätsklinikum Düsseldorf, Düsseldorf, Germany

Csilla Oláh
Department of Urology, University Hospital Essen, University of Duisburg-Essen, Hufelandstraße 55, D-45147 Essen, Germany

Anita Csizmarik
Department of Urology, Semmelweis University, Ulloi ut 78/b, 1082 Budapest, Hungary

Boris Hadaschik
Department of Urology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Fabinshy Thangarajah
Department of Gynecology and Obstetrics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany

Laura Reetz
Department of Pediatrics III, University Hospital Essen, University of Duisburg-Essen, 45147, Essen, Germany

Jamal Ghanam
Department of Pediatrics III, University Hospital Essen, University of Duisburg-Essen, 45147, Essen, Germany

Basant Thakur
Department of Pediatrics III, University Hospital Essen, University of Duisburg-Essen, 45147, Essen, Germany

Halime Kalkavan
University Duisburg-Essen  https://orcid.org/0000-0002-0090-4080

Martin Schuler
University Hospital Essen  https://orcid.org/0000-0002-2166-3394

Jens Siveke
Bridge Institute of Experimental Tumor Therapy, West German Cancer Center, University Hospital Essen  https://orcid.org/0000-0002-8772-4778

Alexander Schramm
University of Duisburg-Essen  https://orcid.org/0000-0001-7670-7529

Alexander Rösch
Keywords:

Posted Date: December 15th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3576016/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Yes there is potential Competing Interest. BH: Advisory boards for Janssen, Bayer, ABX, Lightpoint, Amgen, MSD, Pfizer, Novartis. Invited speaker for Accord, Astellas, Janssen R&D. Honoraria from Uromed. Research funding from AAA/Novartis, Bristol Myers Squibb, and German Research Foundation. Leadership roles for DKG AUO and DGU. MM reports outside the submitted work honoraria for advisory boards from Amgen, AstraZeneca, BMS, GSK, Janssen, MSD, Novartis, Novocure, Pfizer, Sanofi, Takeda MS reports Potential Conflicts of interest • Consultant (compensated) – Amgen, AstraZeneca, Blueprint Medicines, Boehringer Ingelheim, Bristol Myers Squibb, GlaxoSmithKline, Janssen, Merck Serono, Novartis, Roche, Sanofi, Takeda • Stock ownership – none • Honoraries for CME presentations – Amgen, Boehringer Ingelheim, Bristol Myers Squibb, Janssen, MSD, Novartis, Roche, Sanofi • Research funding to institution – AstraZeneca, Bristol Myers Squibb J.T.S. receives honoraria as consultant or for continuing medical education presentations from AstraZeneca, Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, Immunocore, MSD Sharp Dohme, Novartis, Roche/Genentech, and Servier. His institution receives research funding from Abalos Therapeutics, Boehringer Ingelheim, Bristol-Myers Squibb, Celgene, Eisbach Bio, and Roche/Genentech; he holds ownership and serves on the Board of Directors of Pharma15 All conflicts of interest are out of the submitted work
Abstract

DNA methylation is an extensively studied, stable, and fundamental epigenetic alteration in most cancer types\textsuperscript{1}. Single-base-pair resolution analyses of DNA methylation is currently feasible\textsuperscript{2}. Analysis of DNA methylation, in liquid biopsies hold practice-changing potentials\textsuperscript{3-6}. Despite undeniable progress, clinical translation lags behind, mainly due to:

1) Challenges associated with DNA methylation analysis.

2) Fragmentation of circulating cell-free DNA (ccfDNA), worsened by bisulfite treatment.

3) Lack of clinical validation for reported ccfDNA methylation markers.

4) Limited functional characterization of ccfDNA methylation markers in tumors\textsuperscript{7}.

We addressed these challenges by creating a comprehensive pan-cancer cfDNA methylation resource, utilizing pools comprising over 140 patient samples and assess the utility of this resource in over 500 patient plasma and tissue samples spanning around 15 cancer entities with different clinical phenotypes and treatment approaches. Furthermore, we developed a pan-cancer enzymatic digital PCR approach and optimized entity-specific assays for ccfDNA methylation scoring. We demonstrated that this resource can profile methylation in unexplored entities, and ccfDNA methylation patterns align with those in tumor samples. Additionally, we unveiled unconventional epigenetic regulation by methylated DNA-binding transcription factors, with tissue- and context-specific and dosage-dependent activities. This work provides a reference resource for identifying minimally invasive epigenetic markers and opens avenues for characterizing methylated DNA-binding transcription factors.

Main

The molecular analysis of plasma-derived circulating cell-free DNA (ccfDNA) for cancer management, commonly referred to as 'liquid biopsies,' has become very prominent over the past decade\textsuperscript{8,9}. Interrogating tumor-derived ccfDNA enables the prediction of early relapse and the evaluation of tumor molecular architecture and evolution\textsuperscript{10-12}. Regrettably, the molecular profiling of liquid biopsies often focuses on identifying hotspot mutations in well-known, cancer-associated genomic regions, which are only detectable in a small subset of patients. This limitation significantly restricts the scope and clinical potential of liquid biopsies, particularly for early tumor detection\textsuperscript{13-17}. Epigenetic alterations, especially DNA methylation, are specific to tissue and cancer types, and they do not face the same limitations\textsuperscript{18}. Expanding the molecular analysis of liquid biopsies to incorporate epigenetic signatures can extend its scope and, consequently, enhance its clinical utility, particularly in early disease stages\textsuperscript{17}.

Recent studies have successfully utilized ccfDNA-derived epigenetic signatures for the detection of various malignancies\textsuperscript{19,20}. Most notably, these epigenetic signatures have demonstrated promise in cancer detection and determining its tissue of origin (TOO)\textsuperscript{21-24}, representing a valuable asset in
identifying the primary tumor in cancers of unknown primary (CUP). Another potentially practice-changing contribution of liquid biopsy markers in clinical oncology is the ability to detect neoplastic diseases before any clinical signs of cancer appear. This can occur both at the initial diagnosis and after curative therapy, allowing for the anticipation of recurrence. Using a straightforward blood test to identify hidden cancers in asymptomatic individuals at a stage when the disease is still treatable holds the potential to revolutionize clinical oncology.

Although entity-specific screening tests are in use, they continue to face the challenge of high false-positive rates\(^{25-28}\). A multi-cancer screening test for early detection has been launched, with several others currently in development\(^ {29\,13,30} \). Each test focuses on a different liquid biopsy compartment or analyte, but the majority of these tests are primarily designed for disease detection, lacking the ability to reveal the tissue of origin (TOO), predict treatment responses, or stratify patients. On the other hand, ccfDNA methylation-based tests show great promise. Nevertheless, they face persistent challenges, notably the constant need for bisulfite treatment of highly fragmented ccfDNA, as well as the complexities involved in developing specific assays for bisulfite-treated DNA. Additionally, there are unresolved uncertainties regarding the tumor relevance of ccfDNA methylation markers.

To the best of our knowledge, a multi-purpose resource of tumor-relevant pan-cancer ccfDNA methylation marks, evaluated under various treatment modalities in a non-invasive liquid biopsy setting, is currently lacking. Furthermore, a comprehensive comparative analysis of ccfDNA methylation marks in tumors is essential to strengthen the clinical relevance of these marks and broaden their application as biomarkers. Additionally, the functional impact of ccfDNA methylation marks in tumor samples remains largely uncharacterized. Lastly, establishing pan-cancer ccfDNA methylation signatures at the single-base pair resolution, along with the development of robust, widely applicable assays for their examination in laboratories with minimal equipment, is needed. Due to the highly fragmented nature and low plasma concentration of ccfDNA, as well as the challenges associated with designing probe-based assays for bisulfite-treated DNA investigation, the assessment of ccfDNA methylation is still in its early stages.

To address this knowledge gap, we conducted a pan-cancer ccfDNA methylation analysis using a sample pooling strategy. We used whole-genome bisulfite sequencing (WGBS) to establish a versatile ccfDNA methylation panel across eleven different cancer entities. We further use reduced representation bisulfite sequencing (RRBS) to validate this panel on individual patient plasma and tissue samples from 11 cancer entities, including two additional cancer entities not previously included in the WGBS pooling approach as well as CUP. We demonstrate the consistency of DNA methylation marks between patient ccfDNA and tumor tissue samples using pan-cancer data of 15 cancer entities from The Cancer Genome Atlas (TCGA) and in-house generated tumor DNA methylation data sets. We further employed an in-house digital polymerase chain reaction (dPCR) assay, based on methylation-sensitive restriction enzymes, to test pan-cancer and entity-specific ccfDNA methylation marks in larger patient cohorts, including two additional cancer entities, validating the panel and identifying consistent markers across different analytical platforms. This panel proved applicable for predicting immunotherapy response, patient stratification, and cancer subtyping. Lastly, we analyzed tissue DNA methylation from fifteen cancer
entities in TCGA and integrated tissue and ccfDNA methylation data to independently validate and functionally characterize our ccfDNA methylation panel. In conclusion, we've established a versatile ccfDNA methylation panel, serving as a valuable resource for functional studies and biomarker evaluation in larger patient cohorts, with potential applications in clinical practice.

Results

Establishment of a pan-cancer ccfDNA methylation panel

We embarked on the creation of a comprehensive resource for cancer-specific ccfDNA methylation sites across diverse tumor entities, encompassing a wide range of clinical phenotypes, including distinct disease stages, responders and non-responders to various therapies, tumor molecular phenotypes, and disease stages. We conducted Whole Genome Bisulfite Sequencing (WGBS) with 100 bp paired-end long reads at a mean depth of 30x (ranging from ~ 6x to 40x). This was performed on pooled samples of ccfDNA and buffy coat-derived DNA from 141 cancer patients and 24 healthy blood donors. The patient group consisted of comparable distribution of males and females, aged 30 to 90 years, with both localized and metastatic disease. The WGBS pools covered the following cancer entities, including bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), skin cutaneous melanoma (SKCM), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), lung adenocarcinoma (LUAD), sarcoma (SARC), medulloblastoma (MB), uveal melanoma (UVM), and CUP. For CUP, we included samples from eight patients, while each other tumor entity was represented by 12 patients, each with a minimum of two samples (a treatment-naïve sample and one under treatment). Our cohort incorporated samples from both responders and non-responders to systemic therapies, including molecularly targeted therapy, chemotherapy, and immune checkpoint blockade. We analyzed ccfDNA samples from 12 healthy female and 12 healthy male blood donors aged between 18 and 61 as controls. Additionally, buffy coat-derived DNA was obtained from patients with head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and CUP with 15 patients in each group. Buffy coat samples were included because we hoped, hoped for epigenetic signatures from blood cells that could be relevant for immune checkpoint inhibitors. We also included buffy coat-derived DNA from the same 24 healthy blood donors. From each sample type, we generated three DNA pools, resulting in a total of 4 sets of 3 DNA pools covering all the analyzed entities. A detailed representation of the sample pooling strategy is available in extended data table 1.

Enzymatically converted WGBS libraries were created from all the pools and sequenced on two lanes of a NovaSeq 600 (as depicted in Fig. 1a). For the establishment of a pan-cancer methylation panel, we performed a differential methylation analysis by comparing patient samples pools to healthy donor samples. Methylation sites with a mean methylation difference greater than 0.1 were selected for panel design. It is worth noting that a higher threshold of > 0.2 is typically used for ccfDNA in an entity-specific setting^6, but we opted for a lower threshold to account for variations in methylation levels among different cancer entities. This resulted in a panel covering approximately 1Mb with probes targeting both methylated and unmethylated sites. Subsequently, we tested the panel on individual ccfDNA and tissue
DNA samples from multiple cancer entities using a hybrid-capture RRBS approach, as illustrated in Fig. 1b. This approach enabled the identification of both entity-specific and pan-cancer ccfDNA methylation markers. We validated individual markers using digital PCR (dPCR) in a larger collection of pan-cancer ccfDNA samples (N = 115 cases and 81 healthy donors), lung adenocarcinoma tissue (N = 60 patients and 22 adjacent), and ccfDNA from lung adenocarcinoma (83 cases and 24 healthy donors), as shown in Fig. 1c.

We next aimed to assess technical differences both within and between sample pools and determine the most informative sample type. Irrespective of donor health status, sample pools of the same type (plasma or buffy coat) exhibited similar coverage, with most achieving a mean sequencing depth of 20x (as shown in Fig. 1d and Extended data Fig. 1a). We observed comparable methylation density across all individual sample pools (Fig. 1d, lower panel), and they clustered based on donor status and sample type (Fig. 1e and Extended data Fig. 1b). Notably, ccfDNA pools from patients formed a cluster, but were more dispersed compared to healthy donor ccfDNA pools (Fig. 1e and Extended data Fig. 1c) similarly, buffy coat-deried DNA from distinct clusters based on donor health status, although these later clusters were very close to each other. Consistent with these findings, there were fewer differentially methylated regions (DMRs) in buffy coat samples compared to ccfDNA samples (Fig. 1f). Further analysis revealed a significantly higher fraction (p = 0.0164) of immune cell-derived DNA in ccfDNA from healthy donors in comparison to patient samples. In contrast, no substantial differences were observed in immune cell-derived DNA contribution between buffy coat-derived DNA from patients and healthy donors (Fig. 1g). Subsequent analyses, therefore, focused on ccfDNA and tissue DNA samples.

**Assessment of the ccfDNA methylation marks in tumors**

Having established a candidate ccfDNA methylation panel, we aimed to assess the clinical relevance of observed disease-associated ccfDNA methylation features. We conducted differential methylation analyses using human 450k methylation data, comparing tumor to non-tumor tissue in 15 cancer entities from TCGA datasets. Each entity was analyzed independently, and methylation sites meeting the criteria (FDR < 0.05 and mean methylation difference > 0.1) were retained (see extended data table 2). Sites meeting these criteria across all 15 entities resulted in 76,042 sites. From WGBS of ccfDNA, we retained 278,936 differentially methylated positions (DMPs) with mean methylation differences > 0.1. Comparing both DMP lists (TCGA pan-cancer and ccfDNA WGBS), we found 1,454 overlapping sites (see Fig. 2a and extended data table 3). A regression model based on these 1,454 sites, when applied to the TCGA data, demonstrated excellent diagnostic performance with an area under the curve (AUC) of 0.99 (Fig. 2b). To reduce the number of DMPs to a manageable level, we focused on the three best-performing features associated with tumors (Fig. 2c, red dots and extended data table 4). These sites were all hypermethylated when projected onto WGBS data from patient ccfDNA pools (Fig. 2d). Aberrant methylation of these selected sites was prognostic, and no apparent methylation differences were observed between different disease stages (Fig. 2e-g). Manual verification of the selected DMPs sites confirmed clear hypermethylation in tumor samples across all analyzed entities (Fig. 2h and Extended data Fig. 2a&b). We validated the methylation status of the three DMPs in tumor vs. non-tumor tissue for
two randomly selected entities, LUAD and PAAD, using data from the Gene Expression Omnibus GSE66836 for LUAD and GSE49149 for PAAD. In these publicly available datasets, we observed hypermethylation in patient samples (see Extended data Fig. 2c). These findings underscore the disease relevance of the DMPs included in our pan-cancer panel, justifying further investigation.

**Analysis of plasma and tissue samples by RRBS**

To evaluate the robustness and versatility of the established ccfDNA panel, we used RRBS (150 bp paired-end). We applied the panel to over 150 patients, including patient ccfDNA and tumor DNA samples as well as non-tumor tissue and healthy blood donor ccfDNA samples. Two cancer entities (uterine corpus endometrial carcinoma (UCEC) and HNSC) that had not been included in the exploratory WGBS were included at this stage. The ccfDNA samples analyzed comprised n = 16 LUAD (eight with localized disease and eight with metastatic disease), n = 8 SKCM (metastatic), n = 8 UCEC (localized), n = 8 HNSC (localized), n = 8 COAD (metastatic), n = 8 CUP, n = 18 PRAD (localized & metastatic) n = 9 BLCA (metastatic), n = 8 PAAD (metastatic), n = 12 BRCA (localized), n = 7 UVM (metastatic), n = 2 SARC (metastatic), and n = 8 healthy donors. For tissue samples analyzed, samples included n = 8 PRAD tumors, n = 4 PAAD tumors, n = 8 LUAD tumors, n = 4 non-tumor lung tissue, n = 8 BRCA tumors and n = 4 non-tumor breast tissue. We determined differentially methylated ccfDNA sites for each entity by comparing ccfDNA from patients vs healthy donors. DMPs were defined as sites with at least 10x coverage, a mean methylation difference ≥ 0.1 and a p-value for differential methylation < 0.05. For entities with tissue data, the DMPs were projected onto the tissue data. Similar to the TCGA data (see above), ccfDNA DMPs reflected the DNA methylation pattern in tumor tissue in all tested entities (Fig. 3a, Extended data Fig. 3a). For LUAD, where we analyzed ccfDNA from localized and metastatic disease, LUAD-specific hypermethylated sites were identified irrespective of disease stage (Fig. 3a, right panel). Some of these sites were more methylated in advanced compared to localized disease (Fig. 3b). All sites were however hypermethylated in patients compared to healthy donors. Furthermore, RRBS analysis of patient ccfDNA revealed differences in ccfDNA methylation patterns between responders and non-responders to immune checkpoint inhibitors in four cancer entities (LUAD, SKCM, BLCA and COAD); this demonstrates that the heterogeneity in the patient population used in the discovery cohort could be recapitulated (Extended data Fig. 3b). Interrogating the versatility of our panel in an entity that was not included in the discovery cohort, we observed distinctive ccfDNA methylation patterns in HNSC patient vs. healthy donor samples, and in HPV-driven vs. non HPV-driven HNSC (Fig. 3c); this finding indicates robustness and versatility of the panel. In the absence of data from corresponding HNSC tumor tissue that could support these observations, we analyzed DNA DMPs present on the human 450k array in the TCGA HNSC cohort. In multivariate and univariate analyses, most ccfDNA DMPs were prognostic in the TCGA dataset (Fig. 3d&e), and most hypermethylated sites in ccfDNA of HNSC patients were likewise hypermethylated in the TCGA tumor tissue data (Fig. 3f&g). These findings support the applicability of this panel of DMPs for untested cancer entities and other clinically relevant phenotypes for minimally invasive biomarker discovery.

**Investigating the discriminatory capacity of the panel**
To confirm that the panel's robustness and versatility were not due to random signals, we evaluated its discriminatory capacity. First, we assessed the identification of entity-specific methylation sites by iteratively analyzing the differential methylation of each cancer entity (excluding CUP) against all others, using therapy-naïve samples only. We identified entity-specific DMPs, with varying numbers observed across different entities (Fig. 4a). Second, we assessed whether DMPs distinguishing responders and non-responders were universal or specific to treatment modality. This involved comparing DMPs between responders and non-responders to immune checkpoint blockade with those between responders and non-responders to different systemic therapies in metastatic BLCA. For these analyses, we generated five patient pools, each consisting of 5 individual patients, for immune checkpoint inhibitor therapy, and four patient pools for platinum-based chemotherapy. We identified distinct response-associated DMPs for platinum and immune checkpoint inhibitor therapy, with less than a 5% overlap (Fig. 4b&c). Third, we showed that DMPs in BRCA differed between the breast cancer subtypes (Fig. 4d). In LUAD, methylation events associated with checkpoint inhibitor response clustered in genomic blocks, primarily on chromosome 11. In contrast, DMPs predictive of response to immune checkpoint inhibitors in other cancer entities were distributed across various genomic regions (Fig. 4e). These findings support the robustness and versatility of our panel, highlighting its high selectivity and specificity.

**Development of targeted methylation assays**

CcfDNA degradation and the challenges associated with DNA methylation analysis, particularly using sequencing and PCR-based approaches, can complicate clinical implementation. To address this, we developed an alternative DNA methylation scoring approach that is potentially more straightforward than standard bisulfite conversion-based methods. Using our RRBS data, we identified pan-cancer DMPs in ccfDNA that have a restriction site in a targeted CpG locus for the methylation-sensitive restriction enzyme HAP II (Fig. 5a). Using 115 ccfDNA samples from patients across 7 cancer entities (including two new entities: ovarian cancer (OV) and renal cell carcinoma (RCC), along with 81 healthy blood donor samples, we conducted probe-based dPCR on bisulfite-converted ccfDNA as the reference standard. The targeted CpG site was hypermethylated in all samples except for BRCA (Fig. 5b). Evaluating the diagnostic performance of this approach yielded an AUC of 73.5% (95% CI: 65.7%-81.4%) as shown in Fig. 5c. Subsequently, we established a dPCR assay to score methylation at the same target CpG site using methylation-sensitive restriction endonucleases. Linearity and limit of detection were assessed by using artificially methylated and unmethylated DNA via EvaGreen®-based dPCR, showing a linear increase in detected DNA molecules for methylated DNA after digestion with HAP II, while unmethylated DNA yielded only background signal (Fig. 5d). The dPCR assay displayed linear performance down to 0.4 ng of input DNA, which is readily achievable with ccfDNA. To normalize detected DNA methylation events, we utilized two methods: normalization to the input DNA amount and normalization to the total number of DNA copies detected in parallel undigested samples. Both methods exhibited equal performance, with an R value of 0.7 and a p-value of 3e-4. (Fig. 5e). We assessed the performance of an entity-specific DMP (localized LUAD) and a pan-cancer DMP using this enzymatic dPCR approach on fresh-frozen tumor and non-tumor lung tissue. Both assays demonstrated significant hypermethylation in tumor samples (Fig. 5f, left panel). Similar results were observed when analyzing ccfDNA from 83 stage IV LUAD patients and 24
healthy blood donors; the pan-cancer DMP marker exhibited higher methylation counts than the entity-specific marker, as anticipated (Fig. 5f, right panel). Since most clinically available tumor samples are formalin-fixed paraffin-embedded (FFPE), we tested the enzymatic dPCR assay in a larger cohort (n = 45) of DNA samples from FFPE tissue of localized LUAD patients. It exhibited similar performance to fresh-frozen samples (Fig. 5g). Furthermore, we assessed the enzymatic dPCR approach using an entity-specific DMP on LUAD ccfDNA samples, achieving a diagnostic performance of 94.4% (CI 89.4%-99.4%) (Fig. 5h). To make sure that the LUAD-specific site was indeed not a broad cancer marker, we analyzed its methylation status in publicly available COAD data (Fig. 5i) and in cdNDA from several cancer entities using methylation sensitive restriction enzyme (Fig. 5j). In both scenarios, the site was hypermethylated in LUAD. With the entity-specific dPCR assay successfully working on the LUAD-specific site, we extended the testing to the pan-cancer DMP marker using the enzymatic dPCR approach across different entities. The assay produced similar results with a slightly lower methylation count for SKCM (Fig. 5l), but overall yielded a higher AUC (Fig. 5k). The pan-cancer methylation profile of the analyzed pan-cancer site is presented in Fig. 5m. These findings highlight the feasibility of scoring DNA methylation from ccfDNA using cost-effective platforms, potentially expanding the utility of liquid biopsies in biomarker research and clinical applications.

Functional implications of ccfDNA methylation marks in vivo

To assess the mechanistic relevance of the ccfDNA methylation panel for solid tumors, we examined whether ccfDNA DMPs were linked to gene expression regulation in vivo. We conducted Enhancer Linking by Methylation/Expression Relationships (ELMER) analysis on the 15 cancer entities from the TCGA mentioned earlier. To streamline the analysis, we raised the differential methylation cut-off to 0.3. Figure 6a displays the methylation pattern of all pan-cancer DNA hypermethylated sites in this analysis. We conducted feature selection as described earlier and identified nine promising candidates, all of which displayed strong associations with disease status (Fig. 6b). Among these, three (cg12018098, cg26104752, cg26509691) were annotated to protein-coding genes (*PITX1*, *SPRED3*, and *TBX15*, respectively) (extended data table 5). Our focus was on *PITX1*, primarily due to its highest expression in tumors (extended data Fig. 4a & extended data table 6) and its known regulation through methylation in various cancer types. The DMP cg26509691 is located within a CpG island on chromosome 5 around the 3'-UTR of the *PITX1* gene. Our ELMER analysis revealed an inverse correlation between cg26509691 methylation and *PITX1* gene expression in certain cancer entities, including HNSC and CESC (extended data table 5). Cg26509691 displayed hypermethylation in all tested cancer entities (Fig. 6c), and *PITX1* gene expression showed aberrations in most analyzed entities, with some showing downregulation and others showing upregulation (Fig. 6d). Conversely, while the DMP cg12018098 is also situated within a CpG island in the gene body of the *SPRED3* gene and exhibits hypermethylation in all analyzed tumors (extended data Fig. 4b), no significant differences in *SPRED3* gene expression were observed, as in the case of *PITX1* (extended data Fig. 4c). Further investigating the regulatory interaction between cg26509691 and *PITX1*, we made an intriguing discovery. While a negative correlation was observed between the methylation level at cg26509691 and the expression of *PITX1* in CESC (Fig. 7a), a positive
correlation was observed for the same gene probe-pair in PAAD (Fig. 7b and extended data Fig. 5a left and middle panels).

Out of the 20 nearby genes in the vicinity of cg26509691, only PITX1 and the antisense transcript C5orf66-AS1 appeared to be influenced by methylation at cg26509691, and the correlation was consistently in the same direction across all entities for both genes (Fig. 7a &b, extended data Fig. 5b). Methylation of both genes has been previously recognized as a prognostic marker in HNSC\(^ {37}\). Notably, these genes are located adjacent to each other on chromosome 5 in close proximity to cg26509691 (Fig. 7c). To understand the tissue-specific correlation patterns between methylation and gene expression, we initially considered the possibility that both activating and repressive transcription factors (TF) might share the same binding site in an entity-specific manner. To explore this, we identified all master regulatory TFs significantly associated with this probe-gene pair (Fig. 7d) and examined their expression in CESC and PAAD. Among these TFs, E2F7, a repressive member of the E2F TF family, exhibited high expression in CESC and lower expression in PAAD (extended data Fig. 5a, right panel). E2F7 belongs to the E2F transcription factor family and is known as a transcriptional repressor\(^ {38,39,40}\). To verify that these findings were not coincidental, we examined the expression of E2F7 in the other thirteen cancer entities and identified entity-specific correlation patterns for the same probe-gene pair (extended data Fig. 5b). Generally, E2F7 expression was higher in entities exhibiting an inverse correlation between cg26509691 methylation and PITX1 expression, with the exception of BLCA (extended data Fig. 5c).

Published data suggest that PITX1 gene expression may be regulated by E2F7 and other transcription factors, including estrogen receptor-alpha (ER\(\alpha\))\(^ {41}\), TFDP1 and E2F2\(^ {42}\). We investigated whether the activity of these TFs was regulated by methylation patterns in a context-specific manner. To address this, we conducted correlation analyses between all TFs associated with the PITX1-cg26509691 gene-probe pair and several genes surrounding cg26509691. Certain TFs, such as E2F2 and KLF5, exhibited positive correlations with PITX1 gene expression, while other TFs were negatively correlated with gene expression in both tumors and non-tumor samples (e.g., NR1H4), or displayed context-specific correlation patterns (e.g., WT1 and PROX1) (Fig. 7e&f). These findings prompted us to analyze all 20 genes surrounding cg26509691 concerning their correlation with the master regulators. Notably, the correlation patterns observed in non-tumor tissue were not entirely replicated in tumor tissue (extended data Fig. 6a&b). Surprisingly, correlation patterns showed significant heterogeneity even among different tumor entities (extended data Fig. 6c). These findings, together suggest, that the tissue expression patterns of different TFs constitute significant variables, that might need to be considered to better understand the different paradigms of epigenetic regulation of gene expression control.

Subsequently, we investigated whether variations in the expression patterns of these TFs within different tissues could account for their regulatory activity. Remarkably, unsupervised clustering of all tumor and non-tumor samples from 15 different entities, based on the expression of these TFs, resulted in distinct clustering patterns for samples with the same histology, depending on whether they originated from diseased or normal tissue (Fig. 7g). While the expression patterns of KLF5 and E2F2 exhibited similar trends to that of PITX1 across tumors, these patterns were not observed in non-tumor tissue (Fig. 7h).
Finally, we examined the expression of *SPRED3, PITX1*, and *TBX15* in other independent cohorts and found similar expression patterns (extended data Fig. 7). In summary, these findings suggest that DMPs defined in ccfDNA may be linked to the regulation of gene expression in tissue. Importantly, these findings challenge the traditional belief in the inverse relationship between gene expression and DNA methylation. It introduces a regulatory model where expression levels must be understood in context, with the tissue of origin being a crucial covariate to consider.

**Predicted transcription factors bind methylated DNA**

We proceeded to investigate whether these predicted TFs, the majority of which are not pioneer factors, exhibit binding to methylated DNA. To do so, we retrieved TF ChIP-seq data from the ENCODE database (https://www.encodeproject.org/chip-seq-matrix) for the following master regulators associated with *PITX1*: *E2F2, KLF5, KLF15*, and *TFDP1*. We included *MBD2* as a positive control for a methylated DNA binding protein. Subsequently, we assessed the peak locations obtained from all these TFs in relation to the transcription start site (TSS). As anticipated, the most robust peak was observed for *MBD2*, followed by *TFDP1*. Intriguingly, all TFs exhibited complete overlap of their peaks, with each of them centered on the transcription start site (TSS) (Fig. 7i). Furthermore, a significant overlap was evident in the genes regulated by these TFs (Fig. 8a). Most notably, the TFs exhibited peaks around the promoter and gene body of *PITX1*, with *TFDP1* showing a particularly strong association (Fig. 8b). Considering that gene promoters are typically rich in CpG regions, we examined the genomic location of the peaks for each of the TFs. We found that a substantial proportion of peaks (ranging from 30–50%) were situated within 3000 base pairs around the gene promoter (Fig. 8c). Additionally, the binding motifs of each of these TFs contained at least one CpG site (Fig. 8d). The TFs identified through the ELMER analysis are therefore capable of binding methylated DNA and influencing gene expression. A comprehensive list of all TFs associated with hypermethylated DNA across all 15 cancer entities is presented in extended table 5.

Finally, we aimed to gain functional insights into the molecular functions of *PITX1*. To do this, we compared samples with *PITX1* expression in the 75th percentile to those in the 25th percentile. We observed differential expression of several HOX genes and conjugating enzymes in both groups (Fig. 8e). Moreover, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis indicated a strong association between high *PITX1* expression and cellular metabolic activities (Fig. 8f).

Taken together, these findings reveal potential novel methylated DNA-binding TFs and further suggest that some of these TFs might play a role in regulating *PITX* gene expression, highlighting the potential impact of methylated DNA-binding TFs on gene expression control. Furthermore, the data suggests a possible role for *PITX1* in the regulation of energy and xenobiotic metabolism.

**Discussion**

Early detection of cancer is critical for improving patient outcomes, as it allows for timely and potentially more effective treatment interventions. Developing and implementing accurate and accessible early detection methods remains a top priority in cancer research and clinical practice\(^43\). Late detection of
treatment failure can indeed have serious consequences, including unnecessary exposure to toxic treatments, increased healthcare costs, and reduced quality of life for patients. Regular follow-up assessments and the development of more sensitive and specific monitoring tools can help identify treatment failure early, allowing for adjustments in treatment plans and potentially improving overall patient outcomes. The use of ccfDNA methylomes for sensitive and early detection of malignancies has been reported for a few cancer entities. CcfDNA methylation can be used to infer tissue of origin and methylation atlases of different human cell types have been established. Despite these indisputable successes in characterizing ccfDNA methylomes in malignancies, clinically available diagnostic tools based on the analysis of ccfDNA methylation for early diagnosis of otherwise asymptomatic patients are still lacking. At least in part, this is due to the challenges associated with the analysis of ccfDNA methylation, the lack of readily available tools for assay development, the limited knowledge of the disease relevance of reported DMP markers and, most importantly, the absence of cross-entity comparisons for identification of both entity agnostic and tumor-specific methylation markers.

Here, we leverage an approach for such development by pooling ccfDNA samples from multiple tumor entities, clinical phenotypes and sampling time points and deploy the strengths of ultra-deep sequencing with tissue-specific epigenetic imprints to generate a pan-cancer ccfDNA methylation resource, which allows identification of generic and entity-specific methylation markers. Our study provides compelling evidence that the methylation patterns observed in circulating cell-free DNA (ccfDNA) faithfully mirror the methylation profiles of the source tumors. This likely occurs because actively dividing tumor cells continuously release tumor-related material into the circulation, including ccfDNA containing epigenetic information. The presence of tissue-specific epigenetic fingerprints in ccfDNA further supports this notion. We identify several methylation sites which are consistently hypermethylated in tumor tissue across multiple disease entities. This explains why the pooling strategy results in a highly complex panel, capable of characterizing unanalyzed entities. Tumor cells are in a highly dynamic continuum of several cell states that can adopt different phenotypes upon stimulation (e.g., by therapeutic assault).

Therefore, pooling captures epigenetic imprints associated with different disease phenotypes and, thus allows for tumor classification, as observed in our breast cancer samples. We also show that ccfDNA methylation signatures can stratify patients into responders vs. non-responders to different therapies with minimal overlap between response markers for each therapy, as demonstrated in bladder carcinoma samples. This is also made possible by the complexity of the initial samples included in the discovery phase and suggests the potential of ccfDNA methylation analysis to also aid in therapeutic decision-making.

Liquid biopsy tests can detect cancer in some cases several years before clinical signs and symptoms occur, and single methylation sites in liquid biopsies can explicitly detect cancer. However, such tests have not been developed into routinely applicable tools, partly because of difficulties in assay design for bisulfite-treated DNA and due to challenges in specificity and in processing high throughput DNA methylation data. We have resolved these challenges by providing entity-specific and pan-cancer DMPs
across 15 cancer entities in tumor tissue and ccfDNA. Additionally, we illustrate the strength of methylation-sensitive restriction enzymes in targeted analysis of ccfDNA methylation\textsuperscript{51} and compare their performance to conventional bisulfite treatment. We further demonstrate that enzymatic methylation analysis on fresh frozen tissue is comparable to data from FFPE tissue. This provides a framework for virtually any molecular biologist to engage in DNA methylation analysis using fresh frozen or FFPE tissue as well as ccfDNA.

Lastly, an in-depth understanding of the regulatory interplay between DNA methylation and gene expression allows a more direct interpretation of tumor dynamics and has the potential to inform patient management strategies. We show that this can be performed in real-time by analyzing ccfDNA methylation over time and that ccfDNA methylation marks regulate gene expression in a tissue-, context-, and TF-dependent manner. We deploy an approach that allows identification of master regulatory TFs that bind methylated DNA but are not necessarily pioneer factors\textsuperscript{52}. In summary, we propose a cancer overarching DNA methylation resource (entity agnostic ccfDNA and tumor DNA methylation atlas) for tumor tissue and ccfDNA that can dramatically shape engagement in the field of DNA methylation analysis and accelerate the development of critically needed biomarkers. Further studies are required to determine whether larger patient cohorts and high-risk groups will contribute to the robustness of this resource.

Declarations

Data availability

The DNA methylation data sets analyzed in this study are publicly available via the TCGA or the provided GEO accession number. Raw whole genome and reduced representation cell-free and tissue DNA methylation data generated during this study are available upon reasonable request to the corresponding author. Source data for the mail figures have been provided as extended data tables

Code availability statement

No new code was developed during this study

Acknowledgements

This work was partly funded by the DKTK joint funding initiative granted to the EXLIQUID consortium. Human biological samples and related data were provided by the Westdeutsche Biobank Essen (WBE, University Hospital Essen, University of Duisburg-Essen, Essen, Germany). SSL was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—SI 1549/3-1 and RO 3577/7-1 (KFO 337). RV was supported by Brigitte und Dr. Konstanze Wegener-Stiftung. HK was supported by grants from the Advanced Clinician Scientist Programm UMEA\textsuperscript{2} (Medical Faculty, University Duisburg-Essen), the Federal Ministry of Education and Research 01EO2104 (BMBF) and the German Cancer Aid (DKH 70115382).
Author contributions

MM & GZ recruited patients and participated in study design and manuscript writeup. GLM & EJH performed wet lab analysis. BH, PAH, TS, RV, CK, ITK, SG, SKB, SPL, CK, LB, KB, CD, TH, VG, SB HN, IE, TF, FT, KL and BKT contributed samples. KL, HK and KLP proof read and language edited the manuscript. AS, MS and JTS provided resources and infrastructure. SSL designed the study, performed analyses and wrote manuscript.

Competing interest

All authors have declared no conflict of interest

Additional information

There is no supplementary Information available for this paper. All additional data files are provided ad extended data figures/tables

Correspondence and requests for materials should be addressed to Smiths Lueong (s.lueong@dkfz.de)

References


Methods

Patient samples

Human blood and tissue samples were obtained from diverse sources via collaborative agreements or as part of the DKTK EXLIQUID consortium. Samples were collected either within the framework of clinical trials, clinical translational studies or local biobanking initiatives at the West German Biobank Essen. Tissue samples were collected from resected tumors and adjacent non-tumor tissue. With the exception of 48 lung tumor tissue samples and 8 prostate cancer tissue samples obtained from FFPE sections, all other tissue samples were fresh frozen. Plasma and buffy coat samples from healthy donors were prepared from remnants of blood samples donated at the Department of Transfusion Medicine, University Hospital Essen. Donors were between 18 and 61 years of age and were equally male and female. Plasma samples from healthy donors were processed on the day of blood donation to minimize contamination with germline DNA. The study was approved by the local ethics committee, and all participants provided signed informed consent to use biomaterial for research purposes. The corresponding approval codes are 21-10470-BO, 19-9095-BO, 22-10856-BO, 17-7740-BO, 14-6056-BO, 17-7729-BO, 17-7859-BO, 21-10192-BO, 125/2019, 55-3-2014. The complete set of samples analyzed in this study is described in extended data table 1.

cCFDNA and tissue DNA isolation

Cell-free DNA was isolated from plasma samples using the QIAamp MiniElute ccfDNA kit (cat # 55284, Qiagen, Hilden Germany) following the manufacturer’s instructions. DNA from buffy coats was isolated with the Maxwell® RSC cell DNA kit (cat # AS1400, Promega, Walldorf, Germany). DNA was isolated from fresh frozen tumor and healthy tissues using the Maxwell® RSC Tissue DNA kit (cat # AS1610, Promega, Walldorf, Germany). DNA isolation from FFPE samples was performed using either the Maxwell® RSC DNA FFPE kit (cat # AS1135, Promega) for prostate cancer samples or the QIAamp DNA FFPE tissue kit (cat # 56404, Qiagen) for non-small cell lung cancer samples. All DNA samples were quantified using the QuantiFluor® dsDNA System, (cat # E2670, Promega) and stored at -80°C until further processing.

WGBS

WGBS was performed on pooled samples from healthy donors and cancer patients. Specifically, three ccfDNA pools were created from healthy blood donors and cancer patients, respectively. Likewise, three pools of buffy coat-derived DNA were prepared from patients and healthy donors, respectively. For pooling, equal amounts (5 ng) of DNA from 12 patients were pooled for each cancer entity (except cancers of unknown primaries, for which eight samples were used) and distributed among all three pools. The DNA pools were dried in a speed vacuum and resuspended in 50 µL low-EDTA Tris buffer (10mM Tris, 0.1mM EDTA). For library preparation, 200 ng ccfDNA or covaris sheared buffy coat DNA was used.
Sequencing libraries were prepared using the NEBNext Enzymatic Methyl-seq kit (cat # E7120L, New England Biolabs, Frankfurt am Main, Germany). Libraries were quantified using the QuantiFluor® dsDNA System, (cat # E2670, Promega) and characterized on a High sensitivity D1000 screenTape (cat # 5067-5585, Agilent, Santa Clara, CA 95051, USA). Six unique dual indexed libraries were multiplexed and sequenced on a lane of a NovaSeq 6000 with S4 chemistry (100 bp paired end).

RRBS

A custom DNA methylation panel was established by identifying differentially methylated regions (mean methylation difference of > 0.1) between ccfDNA from cancer pools compared to healthy donor pools and cancer buffy coat vs. healthy donor buffy coat. All differentially methylated regions were concatenated to generate a single bed file for probe design and only the top 5000 regions with “good” quality probes (i.e. probes without predicted off-target binding) as per the manufacturer’s recommendation were included in the final panel. Each region within the panel consisted of at least 10 CpG sites, and 120 bp biotinylated probes were designed to cover all regions of interest. For reduced representation sequencing, at least 15 ng of DNA was used as input into the NEBNext Enzymatic Methyl-seq kit. Unique dual indexed libraries were generated as described above, except that samples from individual cancer entities were not pooled. In addition to the ccfDNA samples analyzed, DNA from tumor and adjacent non-tumor tissue was analyzed according to the same protocol. Regions of interest were enriched by hybrid capture and amplified to generate sequencing-ready libraries. Reduced representation libraries were sequenced on a high-output 150 bp paired-end flow cell of a Nextseq 550, aiming at 100x coverage. All sequencing libraries, including those for the WGBS analysis, were spiked with 10% PhiXv.3 library. A comprehensive description of all samples and corresponding analysis types is given in extended data table1.

WGBS and RRBS computational processing

Fastq files were assessed with the fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) tool and trimmed with trimgalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed reads were mapped to a concatenate version of the human (hg38) and lambda genome using Bowtie 2 as implemented into Bismark. Following deduplication of the resulting mapped reads (only for WGBS), methylation events were extracted using default settings for paired-end data using the Bismark methylation extractor. Downstream analysis of the coverage files was performed with RnBeads53. To this end, the following RnBeads options were set: The genome build was set to hg38, minimum coverage threshold was set to 10, and methylation sites for low coverage regions was set to “missing”. Additionally, methylation events on sex chromosomes, sites with single nucleotide polymorphisms, sites containing NA as methylation value, and sites with low standard deviation for their methylation values were filtered out. Differential methylation analysis was performed using limma54. For RRBS, only sites with a p-value for differential methylation of less than 0.05, a mean methylation difference greater than 0.1, and coverage of at least 10 in the patient samples were considered. Immune cell content was estimated using the leukocytes unmethylation for purity (LUMP) algorithm as implemented in RnBeads53.
Regulatory element landscape and transcription factor network analysis

The ELMER Bioconductor package was used for inferring the regulatory element landscape and transcription factor networks using human 450k beadchip array data from the TCGA. For this analysis, only hypermethylation events in tumors were interrogated to allow integration with data from our panel, which was selected for hypermethylation events only. Only entities with paired DNA methylation and gene expression data from the same samples and with healthy non-tumor controls were included. The list of entities included in the ELMER analysis is provided in extended data table 2. For ELMER analysis, a multi-assay experiment was created including a pooled methylation matrix for all samples from each entity, a pooled gene expression matrix for all samples, a matrix mapping gene expression to DNA methylation, and clinical data matrix. Differential methylation analysis was performed between tumor and non-tumor samples to find hypermethylated probes in tumor tissue. Putative gene probe pairs were identified for each differentially methylated probe. For this purpose, the nearest 10 genes up- and downstream of the differentially methylated site were tested for inverse correlation with default empirical p-value (Pe). Using Hypergeometric Optimization of Motif EnRichment (HOMER), motif enrichment analysis was then performed on all significant probe-gene pairs to find potential regulatory TFs around the probe (± 250 bp) using Fisher’s exact test. All probe sets with an odds ratio greater than 1.3 with and a false discovery rate (FDR) < 0.05 were included. The master regulatory TFs binding to the identified enriched motifs were then identified by comparing the mean DNA methylation at each significant gene-probe pair with the gene expression pattern of all human TFs. For this test, a Mann-Whitney U test is performed for each TF-motif pair; this assumes that coordinated alterations within a group of enhancer regions in a set of samples is a consequence of changes in an upstream master regulatory TF.

ChIP-seq data analysis

Raw fastq files for five transcription factors (E2F2, TFDP1, KLF5, KLF15 and MBD2) were downloaded from the encyclopedia of DNA elements (ENCODE). Read quality was checked with fastqc and mapped to the human reference genome hg38 using bwa. The mapped reads were used to create tag directories and bedGraph files were generated. Peaks were called using TF mode in HOMER and annotated. Known and de novo identified motifs were identified and the data were displayed using trackviewer.

Identification of disease-relevant pan-cancer ccfDNA methylation sites

Differentially hypermethylated sites from each of the 15 cancer entities analyzed in the ELMER procedure were intersected to define a list of cross-cancer tumor-derived hypermethylation events. This list was matched with all differentially hypermethylated events derived from whole genome methylation sequencing of patient ccfDNA. The most relevant candidate sites were prioritized using random forest analysis as implemented in the caret package. To this end, the methylation β-values for all sites identified in the above analysis were downloaded from the TCGA using the TCGAbiolinks package. The β-value matrix was divided into training and test sets (80% and 20%, respectively). Missing data were imputed using KNN imputation. Features with 10-fold cross validation repeated five times were selected by
recursive feature elimination. The model was trained on the processed data and the most important variables were selected.

**Identification of entity-specific ccfDNA methylation sites**

Differentially methylated sites between patient ccfDNA and ccfDNA from healthy blood donors were extracted. Only sites with a mean methylation difference greater than 0.1, a p-value for differential methylation of less than 0.05 and a coverage of at least 10 in the patient samples were considered. For entities for which tumor tissue or tumor and normal tissue was available, the identified differentially methylated sites were projected on the tissue data. Entity-specific ccfDNA methylation sites were determined using data from RRBS. For this purpose, treatment-naive samples from each cancer entity were compared with an equal number of samples from all the other entities pooled into one group. The procedure was repeated for all entities investigated except for cancers of unknown primaries.

**Validation data sets**

For independent validation of pan-cancer and entity-specific markers, 450k tissue methylation data were downloaded from the gene expression omnibus (GEO) for non-small cell lung cancer (LUAD), PAAD and COAD. For pancreatic cancer, the 450k data from the GSE49149 study\(^3\) comprising 196 samples (167 PAAD and 29 adjacent pancreas tissues) from patient between 0.8 to 90 years of age were used. Of the 196 samples in this study, 85 were from females and tumor cell content ranged from 12% to 97%. For LUAD, human 450k methylation data from the GSE66836 study\(^3\) comprising samples from 164 non-LUAD patients (101 females) and 19 samples from matched normal lung were downloaded. Patients included in this study had different disease stages (from stage I to IV). For COAD, 261 samples comprising 149 mucosa and 112 tumor samples from the GSE101764 study\(^5\) were used. Study participants were between 22 and 89 years of age, and included 180 females. All validation datasets were analyzed using the methylationArrayAnalysis Bioconductor package using default settings.

**Targeted analysis of pan-cancer and entity-specific sites by digital PCR**

We developed two digital PCR-based approaches for targeted analysis of cell-free and tissue DNA methylation at single nucleotide resolution. In the first assay, a probe-based bisulfite conversion method was used, while a methylation sensitive restriction enzyme-based (MSRE) approach was implemented for the second one. In the bisulfite conversion approach, DNA samples were converted using the EZ DNA methylation gold kit (cat # D5006, Zymoresearch, Freiburg, Germany). The bisulfite converted DNA was analyzed by probe-based digital PCR with a FAM-labelled probe targeting the methylated event and a HEX probe targeting the non-methylated event. In the MSRE approach, a 400 bp region (200 bp up- and downstream) of the targeted methylation site was extracted and screened for the presence of restriction sites for methylation-sensitive restriction enzymes. Only regions with exactly one restriction site for a given restriction enzyme and lying within the targeted CpG site were further analyzed. Samples were digested with the defined restriction enzyme overnight at 37°C. Following complete digestion, the DNA
was amplified with pre-optimized primer sets and methylation copies were either normalized to the input DNA mass or to copies obtained from the corresponding undigested sample.

**Methods references**


**Figures**
Figure 1

A workflow for the identification of a pan-cancer cell-free and tissue DNA methylation atlas from liquid biopsies. a, A scheme showing the establishment of a targeted DNA methylation panel by whole genome bisulfite sequencing of ccfDNA and buffy coat DNA from pools from cancer patients and healthy donors. Three ccfDNA pools, each containing samples from all studied entities were generated from 12 patients per entity (except for (CUP). Patients were aged between 30 and 90 years with comparable proportions of...
males and females. Samples from MB patients included both ccfDNA as well as DNA isolated from cerebrospinal fluid from several subtypes of brain tumors. CcfDNA from 24 (50% males) healthy blood donors aged 18-61 years was reconstituted into three pools. Buffy coat DNA pools were made from eight patients each from four cancer entities (HNSC, BRCA, NSCLC and CUPs). Healthy control Buffy coat was from the same donors as the ccfDNA. Enzymatic conversion of DNA using TET2/APOBEC system. b, Reduced representation bisulfite sequencing scheme demonstrating the workflow for individual patient tissue and ccfDNA analysis. Between 8-18 individual ccfDNA samples were sequenced per entity plus eight ccfDNA samples from healthy controls. Patients included here were not the same as those used in the discovery cohort. Two additional entities (UCEC and HNSC), which were not present in the discovery phase (WGBS) were included as well as tumor and adjacent normal tissues from four entities (BRCA, NSCLC, PAAD & PRAD). c, A schematic representation of targeted validation of selected CpG sites by means of digital PCR of bisulfite converted and methylation-sensitive restriction enzyme digestion in larger patient cohorts. d, Quality control data presenting the sequencing coverage of ccfDNA and Buffy coat pools from healthy donors and cancer patients (upper panel) and methylation densities between healthy donors and cancer patient samples (lower panel). e, Multidimensional scaling plot showing clustering of donor and sample types from WGBS data. f, a heatmap presenting differentially methylated CpG sites (mean methylation difference ≥ 0.1) between healthy donor and patient ccfDNA and between healthy donor Buffy coat and patient Buffy coat DNA. g, Violin plot showing the predicted immune cell content of ccfDNA between patients and healthy donor and Buffy coat between healthy donor and cancer patients. Immune cell content was estimated using the LUMP algorithm as implemented in the Bioconductor package RnBeads. All images used in a, b & c are made in Biorender. The presented p values are obtained from an unpaired student’s t-test.
Figure 2

Pan-cancer ccfDNA methylation profiles from WGBS mirrors tumor DNA methylation patterns in multiple solid tumors within the TCGA. a, Venn diagram presenting the intersection of differentially hypermethylated positions (mean methylation difference > 0.1 and coverage ≥ 10x) from WGBS pan-cancer ccfDNA pools and differentially hypermethylated positions (DMPs) between tumor tissue and normal tissue obtained from 15 cancer entities within the TCGA (human 450k beadchip arrays). A heat
map showing the methylation patterns of DMPs (> 1400 sites from WGBS data of ccfDNA pools) which are common between pan-cancer WGBS pools and tumor tissue from the TCGA. Differentially hypermethylated sites from the TCGA tissue cohorts were calculated for each entity and cancer overarching sites merged (right panel). b, a receiver operator characteristic curve showing the specificity and sensitivity of a regression model on the common DMPs using tumor/normal tissue data from the TCGA. c, bar plot showing the importance score of the best performing CpG sites in the TCGA pan-cancer data, focusing on sites common in tissue and plasma. The caret package was used for modeling and extraction of importance variable. d, a violin plot on data from WGBS of ccfDNA pools showing the methylation profiles of the top 3 best performing CpG sites observed in the pan-cancer tissue TCGA data in ccfDNA samples from cancer patient and healthy donor pools. e, a forest plot from a multivariate Cox proportional hazards (CoPH) model performed on the top 3 best performing CpG sites. Overall survival estimates for more than 4000 patients (with comprehensive survival data) from the TCGA was included. Presented p values are from a log-rank test. f, a representative Kaplan-Meier overall survival curve showing the overall survival of patients with low (light blue) and high (red) DNA methylation levels for the differentially hypermethylated CpG site cg19416570. Low and high dichotomization was determined using an optimal cut-off as implemented in the survminer package. Analyses were performed on pan-cancer data for fifteen cancer entities of the TCGA. g, violin plots showing the methylation patterns of the three best performing differentially hypermethylated sites as a function of tumor stage in pan-cancer data from the TCGA. h, representative boxplot showing the methylation pattern of the pan-cancer hypermethylated site cg19416570 in tumor vs non-tumor tissue in fifteen cancer entities from the TCGA.
Figure 3

Reduced representation bisulfite sequencing using pan-cancer ccfDNA panel on tumor and non-tumor tissue and previously unanalyzed cancer entities. a, heatmap showing differentially methylated CpG sites (mean coverage \( \geq \) 10, mean methylation difference \( \geq \) 0.1, and p value < 0.05) in ccfDNA from PAAD patients vs healthy blood donors as well as PAAD tumor tissue (left panel) and differentially methylated CpG sites in ccfDNA from LUAD vs healthy blood donor and tumor vs adjacent non-tumor LUAD tissue.
For both PAAD and LUAD, differentially methylated sites were obtained by comparing ccfDNA from patients and healthy donors. The sites were then projected on data from tumor tissue without any further differential analysis. b, boxplots showing methylation profiles for representative differentially methylated sites in LUAD comparing ccfDNA from healthy donors (grey), localized stage LUAD (stage I-III, orange) and metastatic disease (blue). P values are from a one-way ANOVA test. c, heatmap showing the performance of the pan-cancer ccfDNA methylation panel on a previously unanalyzed entity (HNSC) and healthy blood donors (left panel) and a heatmap showing differentially methylated positions between HPV-driven HNSC and HPV unrelated HNSC. For RRBS panel establishment by WGBS, HNSC as well as UCEC were not included. d, a forest plot for multivariate cox proportional hazards model on differentially hypermethylated CpG sites in ccfDNA from HNSC vs healthy blood donors using tissue data from the TCGA. The genomic coordinates of each site were matched on the hg38 assembly and projected to the human 450K beadchip annotation. e, Kaplan-Meier overall survival curves for two significant differentially methylated CpG sites on the HNSC data from the TCGA. f, boxplot showing the methylation patterns of differentially methylated sites on RRBS data from HNSC ccfDNA data. Only sites with annotation on the human 450K beadchip are shown. P values are from a Mann-Whitney-U-Test. g, boxplot showing the methylation β values of the corresponding sites in e, on HNSC tissue data from the TCGA. P values are obtained from a t-test.
Figure 4

**RRBS of patient ccfDNA reveal entity and phenotype-specific methylation patterns.**

**a**, a heat map showing entity-specific methylation sites in ccfDNA samples from eleven cancer entities. Samples from healthy donors and cancers of unknown primaries are excluded from this figure. **b**, heatmaps showing differentially methylated ccfDNA sites in bladder cancer patients who did or did not respond to immunotherapy (left panel) and chemotherapy (right panel). For each treatment, five to six patient
ccfDNA samples were pooled together and analyzed. Each lane on the heatmap represents a patient pool. 

Image c shows a Venn diagram illustrating the overlap between differentially methylated sites in PRAD responders vs non-responders to immunotherapy and responders vs non-responders to chemotherapy. 

Image d displays a principal component analysis in ccfDNA samples from BRCA patients, showing supervised clustering of the different breast cancer subtypes based on ccfDNA methylation profiles. The top 3000 most variable sites were obtained by calculating median absolute deviation and scaling the data thereafter. The scaled data was used for principal component analysis. 

Image e presents boxplots depicting representative methylation profiles for differentially methylated ccfDNA sites in responders vs non-responders to immune checkpoint blockade in SKCM, LUAD, and COAD. All p values are calculated by a Mann-Whitney-U-Test.
Figure 5

Targeted analysis of entity-specific and pan-cancer ccfDNA methylation sites by means of digital PCR. a, a heatmap showing differentially methylated CpG sites across 11 cancer entities vs healthy controls. Each cancer entity is represented by at least 8 independent samples and 8 healthy controls. Differentially methylated sites are defined by a p value < 0.05, mean methylation difference ≥ 0.1 and a coverage of >10. b, boxplot presenting the methylation pattern of a ccfDNA pan-cancer hyper methylated site across...
several cancer entities (N = 115, BRCA = 8, COAD = 12, LUAD = 19, SKCM = 19, OV = 15, RCC = 17, UCEC = 25) and 81 healthy blood donors. Methylation profiles were measured by probe-based digital PCR with dual probes for the methylated and unmethylated allele. c, a receiver operator characteristic curve presenting the diagnostic performance of a pan-cancer ccfDNA hypermethylated site on multiple cancer entities analyzed by digital PCR following sodium bisulfite treatment. d, a line plot displaying the limit of detection and correlation between input DNA mass and number of detected methylated and unmethylated DNA standard after digestion with the methylation-sensitive restriction enzyme HAP II. Methylated and unmethylated DNA standards were purchased from commercial providers and analyzed by digital PCR. e, a scatter plot displaying the correlation in two tested normalization methods for scoring ccfDNA methylation by digital PCR following digestion with methylation-sensitive restriction enzymes. On the y-axis the number of DNA copies detected after digestion is normalized to the input mass of DNA. On the x-axis, the number of detected methylated DNA copies after digestion is normalized to the number of DNA copies detected in a paired undigested sample analyzed simultaneously by dPCR. f, boxplot showing the methylation profile of an entity-specific (LUAD) ccfDNA methylation site and a pan-cancer ccfDNA methylation site on LUAD tumor and non-tumor tissue (left panel) and a boxplot showing the methylation profile of an entity-specific (LUAD) ccfDNA methylation site and a pan-cancer ccfDNA methylation site on ccfDNA from patients with stage IV LUAD (n = 83) vs healthy blood donors (n = 24), right panel. g, boxplot showing the methylation profile of a LUAD-specific ccfDNA methylation site in FFPE derived LUAD tumor tissue (n = 45) and non-tumor lung tissue (n = 22). h, receiver operator characteristic curve presenting the diagnostic performance of an entity-specific (LUAD) ccfDNA hypermethylated site on ccfDNA from 83 LUAD samples and 24 healthy blood donors following digestion with HAP II. i, boxplot of DNA methylation β values showing the methylation pattern of the LUAD specific CpG site, in COAD and LUAD. j, boxplot showing the methylated (undigested fraction) of the LUAD-specific site in cfDNA from multiple entities, including LUAD. k, boxplot presenting the methylation pattern of a pan-cancer ccfDNA hyper methylated site across several cancer entities (N = 81) and 24 healthy blood donors. Methylation profiles were measured by evagreen-based digital PCR and normalized to the undigested control. l, a receiver operator characteristic curve presenting the diagnostic performance of a pan-cancer ccfDNA hypermethylated site on multiple cancer entities analyzed by digital PCR following digestion with HAP II. m, violin plot displaying the methylated fraction of the analyzed pan-cancer site in WGBS data.
Figure 6

Functional analysis of pan-cancer ccfDNA hypermethylated sites with corresponding hypermethylation in tumor tissue. a, heatmap showing the methylation profile of about 320 DNA methylation sites with a mean methylation difference ≥ 0.3 in tumor vs non-tumor tissue across 15 cancer entities within the TCGA. b, bar plot showing the top 9 most distinctive methylation sites across all cancer entities with a mean methylation difference ≥ 0.3 (upper panel) and a table presenting odds ratios, 95% confidence
intervals and corresponding p values for a logistic regression describing the association between each site and tumor disease (lower panel). c, boxplot showing the methylation profile between tumor and normal tissue across 15 cancer entities within the TCGA for a selected CpG sites (cg26509691) with mean methylation difference ≥ 0.3 and which is associated with the protein coding gene PITX1. d, boxplot showing the gene expression profile between tumor and normal tissue across 15 cancer entities within the TCGA for the gene PITX1, which is associated with the DMP cg26509691.
**Figure 7**

ccfDNA identified DMPs regulate gene expression in a context-dependent manner. **a**, scatterplots showing the correlation between tissue DNA methylation at the cg26509691 and the expression of the 20 nearby genes (10 up- and 10 downstream of the CpG site) in cervical squamous cell carcinoma (CESC). **b**, scatterplots showing the correlation between tissue DNA methylation at the cg26509691 and the expression of the 20 nearby genes (10 up- and 10 downstream of the CpG site) in pancreatic ductal adenocarcinoma (PAAD). Reverse correlation patterns are observed in both entities for the gene PITX1 and C5orf66-AS2. **c**, a schematic plot showing the localization the cg26509691 on chromosome 5 and the 20 nearby genes as well as a box and violin plot displaying the methylation pattern of this CpG site in primary tumors, metastatic lesions and non-tumor tissue. **d**, a forest plot showing transcription factors whose binding motifs encompass the sequence including cg26509691 and are significantly associated with the regulation of nearby genes. The odds ratio and 95% confidence interval for the association of each transcription factor is presented as well as the number of probe-gene pairs. **e**, a correlation matrix showing the correlation coefficient between selected cg26509691-nearby genes and all candidate transcription factors associated with probe-gene pairs in on-tumor samples (left panel) and tumor tissue samples (right panel). Some transcription factors are unmatched between tumor and normal because of low or no expression. **f**, a scatter plot showing the correlation between the expression of the transcription factor NR1H4 and the cg26509691-associated gene PITX1 (upper panel) and a scatter plot showing the correlation between the expression of the transcription factor E2F2 and the cg26509691-associated gene PITX1 (lower panel). **g**, umaps presenting the clustering of tumors samples (left panel) and normal samples (right panel) based on the expression of the transcription factors associated with the cg26509691-PITX probe gene pair. **h**, heat maps showing the expression of selected transcription factors in non-tumor samples (upper panel) and in tumor samples (lower panel) across fifteen tumor entities with the TCGA. **i**, a ChIP peak profile plot showing the distribution of ChIP-seq peaks around the transcription start site of genes for five selected cg26509691-associated transcription factors (E2F2, TFDP1, KLF5, KLF15 and MBD2).
Probe-gene pair associated transcription factors bind CpG-rich region and regulate gene expression.  

Figure 8

Venn diagram showing the number of overlapping annotated ChIP-seq peaks between the different transcription factors. MBD2 was purposely included to serve as a control for methylated DNA-binding protein.  

ChIP-seq peak profiles for all five transcription factors surrounding the promoter region of the PITX gene.  

Bar plot showing the distribution of ChIP-seq peak across different genomic regions for all transcription factors.
analyzed transcription factors. d, seqlogo plot showing the JASPER binding motif of analyzed transcription factor. The CpG position is highlighted in red within each binding motif. e, volcano plot showing genes differentially expressed in samples with high transcript levels vs low levels of PITX1. PITX$_{\text{high}}$ samples are those with 75$^{\text{th}}$ percentile and those with low expression are within the 25$^{\text{th}}$ percentile. f, dot plot presenting the enrichment of different KEGG pathways in differentially expressed genes within the PITX$_{\text{high}}$ and PITX$_{\text{low}}$ patient samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- extended datatable1.xlsx
- extended datatable2.xlsx
- extendeddatatable3.xlsx
- Extended datatable4.xlsx
- Extended datatable5.xlsx
- Extended datatable6.xlsx
- Extended data figures.docx