Early signatures of breast cancer up to seven years prior to clinical diagnosis in plasma cell-free DNA methylomes

Nicholas Cheng

Ontario Institute for Cancer Research

Kimberly Skead

Ontario Institute for Cancer Research

Tom Ouellette

University of Toronto

Scott Bratman

Princess Margaret Cancer Centre <https://orcid.org/0000-0001-8610-4908>

Daniel De Carvalho

PMCC

David Soave

Wilfrid Laurier University

Philip Awadalla (✉ Philip.Awadalla@oicr.on.ca)

Ontario Institute for Cancer Research

Article

Keywords:

Posted Date: January 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1203227/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Limited studies to date have investigated the detectability of cell-free DNA (cfDNA) markers in asymptomatic individuals prior to a cancer diagnosis. Here, we performed cfDNA methylation profiling in blood of individuals up to seven years prior to a breast cancer diagnosis in addition to matched cancer-free controls (n=150). We identified cfDNA differentially methylated signatures that discriminated cancer-free controls from pre-diagnosis breast cancer cases over five years prior to diagnosis and demonstrate that these markers were reflective of methylation profiles in breast cancer tissue. We report classification of a range of pre-diagnosis breast cancer cases detected at Stage I (area under the receiver operating characteristic curve (AUC) of 0.771), and in cases with a negative mammogram screening within a year of blood collection (AUC of 0.852). This study provides evidence that cfDNA methylation markers indicative of breast cancers can be detected in blood among asymptomatic individuals prior to clinical detection.

Introduction

High morbidity and mortality rates associated with cancers can largely be attributed to late-stage diagnoses. Across most cancers, survival outcomes are significantly improved when tumours are still localised to the tissue of origin at diagnosis¹. However, effective population screening tools for early cancer detection are currently limited to a few cancer types, notably breast, colorectal, lung and cervical cancer^{2,3}. While mammograms remain the gold standard for breast cancer diagnostic, there are associated economic costs and risks associated with radiation exposure which accompany routine mammogram screening. Further, limited participation, as well as low positive predictive values owing to high false positive rates, have raised concerns of overdiagnosis and overtreatment of breast cancers⁴⁻⁶.

Profiling cell-free DNA (cfDNA) derived from tumours in blood, also known as circulating tumour DNA (ctDNA), is well demonstrated to be a potential non-invasive biomarker as it provides a glimpse into the genetic and epigenetic landscape of a tumour's genome⁷⁻¹⁰. Sensitive liquid biopsy assays examining tumour specific genetic and epigenetic alterations in cfDNA are able to detect both early and late stage cancers and inform the tissue of origin of underlying tumours. In addition, some studies have even combined cfDNA biomarkers with other markers such as multi-protein panels or radiographic imaging to further improve diagnostic accuracy⁷⁻⁹. Several of these early studies, to date, have demonstrated the prognostic potential of using cfDNA methylation profiles for early detection of breast cancer and other cancer types¹⁰⁻¹². However, as the majority of cancers are often detected once patients are screened or become symptomatic, these studies have primarily been performed using biologic samples collected from patients following clinical detection and diagnosis of a malignant primary tumour. Profiling cfDNA in the pre-diagnostic context could allow us to better understand the detectability of cancer biomarkers at the earliest stages, however this requires application of new technologies to biologics collected from healthy individuals prior to a cancer diagnosis.

Here, we profiled cfDNA methylation patterns in plasma samples collected from cohort participants prior to a breast cancer diagnosis, and matched cancer-free controls, to identify cfDNA markers indicative of early breast cancers. We leverage the Ontario Health Study (OHS), an Ontario-based longitudinal prospective cohort that collected health and lifestyle information through self-reported questionnaires, and biologics including blood plasma, from over 41,000 participants between 2009 and 2017 upon initial recruitment to the study¹³. A particular advantage of the OHS is that almost all participants provided consent to administrative health linkages at the time of enrollment into the study. As such, we were able to link the health insurance numbers of recruited individuals to administrative health registries to identify participants that developed breast cancer up to seven years after study recruitment and biologic donation. Using 1.6 mL of blood plasma from participants that developed breast cancer, in addition to matched cancer-free controls, we analyzed and compared cfDNA methylomes in pre-diagnosis blood plasma samples versus controls. In this study, all sequencing runs and analytics were performed with cases and controls concurrently to minimize inflation of accuracy, sensitivity and specificity. By retrospectively interrogating blood samples collected prior to diagnosis, we assessed the earliest detectability and predictive performance of cfDNA methylation markers for classifying participants harboring undetected breast cancers.

Results

In the OHS, we identified 110 participants that provided a blood sample at time of enrollment and developed breast cancer following study enrollment and, in addition to 108 control participants with no history of cancer at the time of study enrollment and throughout the study follow up time. Control samples were matched to each pre-diagnosis cancer case by age, sex, time of sample collection, ethnicity, body mass index, smoking frequency and alcohol consumption frequency (Fig. 1, Extended Data Fig. 1, Table 1 and Extended Data Table 1). Using 1.6 mL of plasma from pre-diagnosis breast cancer and selected cancer-free control participants, we profiled cfDNA methylation patterns using a cell-free methylated DNA immunoprecipitation sequencing protocol (cfMeDIP-Seq), which pulls down and sequences methylated cfDNA fragments^{14,15}. To mitigate confoundment of biological signals from technical artifacts associated with batch effects, all cancer cases were batched together with control samples during and across sequencing runs. Coverage profiles of 300 bp non-overlapping tiled windows across the genome were computed from raw sequencing reads to infer cfDNA genome-wide methylation levels. We applied filters to reduce background signals from healthy tissue cfDNA, as most cfDNA fragments in blood are derived from peripheral blood leukocytes and other hematopoietic cells, (Extended Data Fig. 2)¹⁶. We identified and removed regions frequently methylated in peripheral blood leukocytes (PBLs) using publicly accessible whole-genome bisulfite sequencing (WGBS) methylation data of peripheral blood leukocytes. Additionally, as regions hypermethylated in breast cancers are typically located at CpG dense regions and regulatory sites (Extended Data Fig. 3 and Extended Data Fig. 4), we retained windows with six or more CpG sites and located at CpG islands, shores, shelves, or promoters and enhancers previously identified by the FANTOM5 project, leaving 81,323 windows to interrogate.

To identify cfDNA methylation signatures among the retained regions that can discriminate pre-diagnosis cases from controls, we divided samples from each group into 10-folds. Iteratively, 9-folds (90% of participants) were used as train set samples to identify differentially methylated regions (DMRs) in pre-diagnosis plasma cfDNA between cases and controls. DMRs were identified by fitting negative binomial regression models across retained regions and adjusting for age and batch as covariates. A random forest classifier was built using the top 200 regions hypermethylated among train set breast cancer cases, and then evaluated on held-out samples from the one remaining held-out fold. We performed differential methylation calling and modeling building and assessment iteratively 10 times with a different held-out fold for testing each time. This cross-validation (CV) procedure was repeated 200 times with different fold-splits to mitigate overfitting and enable assessment of the variability in cross-validated predictive performance, effectively performing differential methylation calling, model building and assessing predictive performance 2000 times (Fig. 2) ¹⁷.

Differentially methylated regions in pre-diagnosis breast cancer reflects tumour epigenome

We identified 487 hypermethylated regions recurrently significant across the 2000 repeated CV DMR analyses ($p < 0.05$ across 10% of subsampled DMR calls; Extended Data Fig. 4). Using principal component analysis (PCA), the top 50 most recurrent DMRs are able to partition pre-diagnosis breast cancers from cancer-free controls up to seven years prior (Fig. 3a). Further, to investigate whether the 487 hypermethylated regions identified from pre-diagnosis cfDNA were reflective of breast cancer tissue methylomes and could discriminate breast cancer tissue from other tissue types, we compiled publicly accessible cancer and normal 450k methylation array data from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (Accession number GSE87571 and GSE4286). Among the 487 identified pre-diagnosis cfDNA DMRs, 286 regions spanning 589 CpG sites were probed by 450k methylation array. In total, 194 out of 286 (67.8%) of pre-diagnosis breast cancer DMRs overlap with at least one differentially methylated CpG site (DMCs) observed between TCGA breast cancer vs breast normal tissue, breast cancer vs PBLs, or cancer vs matched normal tissues across 12 tissue types (Extended Data Fig. 5). Additionally, we performed a permutation test to infer whether significant cfDNA DMRs were significantly enriched for, and concordantly methylated, in breast cancer tissue relative to the expected overlap estimated from random subsampling of the genome. Notably, the cfDNA methylation markers were most enriched for CpG islands that are differentially methylated in breast cancer tissues relative to matched normal tissue (Fig. 3b). While age-associated changes in DNA methylation patterns have previously been observed across all tissues and could potentially confound the pre-diagnosis DMRs, we observe no overlaps between the pre-diagnosis DMCs and previously established age-associated markers such as Horvath's epigenetic clock predictors¹⁸. Further, using 86 CpG sites overlapping the top 50 pre-diagnosis breast cancer cfDNA DMRs, we demonstrate with tSNE visualization that plasma-derived pre-diagnosis breast cancer cfDNA DMRs can discriminate breast tissue from normal PBLs (Extended Data

Fig. 6a) and similarly between normal and cancerous breast tissue (Extended Data Fig. 6b). While it was unclear which markers are specific to breast cancers, we observe that the pre-diagnosis breast cancer cfDNA markers can also discriminate between tumour versus normal biopsies in other tissue types such as lung and liver (Extended Data Fig. 7).

Among the 487 hypermethylated regions, we observe significant enrichment for transcription factor binding sites compared to the expected overlap from random subsampling of the genome (Fig. 3c). Interestingly, some of these transcription factors such as *SIN3A* and *RUNX3*, are known to have tumour suppressive functions inferred in previous studies of their deletion and downstream targets^{19,20}. Likewise CTCF binding site methylation is frequently detected across breast and other cancer types²¹. We observe that the binding sites of these tumour suppressors can potentially be disrupted among hypermethylated cfDNA regions. Among proximal gene targets of hypermethylated cfDNA DMRs, various genes with dysregulated methylation are inversely correlated with changes in expression between TCGA breast cancer and normal tissue among overlapping CpG sites (Fig. 3d). Genes of interest include *IRX2* a tumour suppressor disrupted in breast cancers demonstrated to suppress cellular motility and chemokine expression²². Likewise, promoter methylation of genes such as *CDKL2*, have also been highlighted as a cfDNA methylation marker for triple negative breast cancers¹¹. Other notable tumour suppressor genes with promoter hypermethylation identified from pre-diagnosis cfDNA profiles include *BRCA1* and *DNAJC15*^{23,24}. Additionally gene set enrichment analyses for gene ontology molecular functions reveals dysregulation in peptide antigen binding from hypermethylation of HLA regulatory elements, similar to previous reports of higher somatic variant burden and promoter hypermethylation in HLA genes which may facilitate immune evasion of early breast tumours²⁵.

Predictive performance of cfDNA methylation markers

Across the 200 repeated 10-fold CV procedures, we observe consistent classification performance between pre-diagnosis cases versus cancer free controls. We achieve a binary classification area under the receiver operating characteristic curve (AUC) across all breast cancer types, ages and varying pre-diagnosis time intervals of 0.747 (95% CI 0.695 – 0.791) and a mean sensitivity of 41% (95% CI 0.312-0.411) at 99% specificity (Fig. 4a). The AUC and sensitivity of the binary classification performance does not take into consideration the follow-up time of controls, the time to diagnosis among the pre-diagnosis cancer cases, as well as the proportion of true negative cases in the population. Accordingly, we computed a weighted concordance index (C-index): the proportion of pairwise comparisons between all samples in which individuals with the higher risk score develops breast cancer sooner compared to individuals with lower risk scores. Pairs within the calculation were weighted according to age-specific breast cancer incidence for the corresponding follow-up times in the Canadian population (see Methods). Additionally, we calculated weighted time-dependent AUCs (AUC(t), a dynamic measure that calculates

AUC for sample individuals within a given follow-up time²⁶. The AUC(t) was relatively consistent with the binary AUROC, achieving an average C-index of 0.734 (95% CI 0.685-0.781) across the 200 10-fold CV iterations up to seven years prior to diagnosis (Fig. 4b). Likewise, the classifiers perform consistently well among cases diagnosed at stage I, with a mean AUC of 0.771 (95% CI 0.708-0.825; Fig. 4c) and mean C-index of 0.761 (95% CI 0.700-0.823; Fig. 4d). Similarly, the AUC and C-index among controls followed up to at least three years and cases diagnosed over three years following blood collection is 0.723 (95% CI 0.653-0.776) and 0.709 (95% CI 0.645-0.766) respectively, indicative of detectable breast cancer cfDNA methylome markers several years prior to an early-stage diagnosis (Extended Data Fig. 8). Notable differences in predictive performance across different breast cancer subtypes are observed. Hormone receptor (HR) positive breast cancers were the most frequent subtype in this study (n = 47 cases; 82.4%) compared to HR negative (n = 10 cases; 17.6%), similar to frequencies observed in breast cancer cases across the US between 2010-2016 (83.8% HR positive)²⁷. Classification models performed better in identifying early HR positive breast cancer cases in pre-diagnostic blood plasma (n = 47 cases), achieving an average AUC of 0.788 (95% CI 0.732-0.835) and C-index of 0.786 (95% CI 0.720-0.844) (Fig. 4e&f).

Typically, all women between the ages of 50-70 are recommended to receive mammograms biennially in Canada. We investigated whether individuals diagnosed before 50, preceding the age of mammogram eligibility in Ontario, could also benefit from cfDNA tests for early breast cancer detection. When stratifying binary classification performance according to age of diagnosis, individuals diagnosed before 50 (n = 17 cases), achieve high classification accuracy with a mean AUC of 0.898 (95% CI 0.646-1) and a C-index of 0.830 (95% CI 0.574-0.999) (Fig. 4g-h, Extended Data Fig. 9). Furthermore, within our study cohort, 42 cases reported to have a negative breast mammography screen within six months to one year of providing a blood sample to the OHS. When stratifying the classification performance for cases with negative mammogram results within one year of providing blood samples (Fig. 5a&b), our classifiers achieve a mean AUC of 0.781 (95% CI 0.711 – 0.841) and mean C-index of 0.774 (95% CI 0.702 – 0.842). Further, we observe that among these samples with a negative mammogram test, 42.9% are classified as positive for breast cancer using cfDNA methylation signatures at 90% specificity, highlighting that cDNA tests accompanied by mammogram screening may improve sensitivity for detecting early breast cancers (Fig. 5c).

Discussion

Using cfMeDIP-Seq to profile cfDNA methylomes, we were able to capture cfDNA methylation signatures indicative of breast cancer prior to clinical presentation and in some cases before mammogram detection. We highlighted the generalizability of using pre-diagnosis DMRs to classify individuals with underlying breast cancers using a repeated cross-validation strategy and demonstrated that these markers can be potentially detectable up to seven years prior to diagnosis. We found that a majority of pre-diagnosis cfDNA DMRs were concordant with DMCs captured between breast cancer versus PBL and normal breast tissue biopsies, further demonstrating that significant DMRs captured are likely reflective

of breast cancer methylomes. There has been an increasing consensus among recent studies that cfDNA methylation profiles, often combined with other biomarker or imaging-based approaches, can yield the best predictive performance for detecting cancers at early stages. However, to implement liquid biopsies for population screening of cancers, the viability of existing assays and predictive models needs to be demonstrated in biologics collected prior to a cancer diagnosis. Our work builds on major investments into developing large longitudinal population cohorts that store biologics collected from healthy individuals at the time of study recruitment. By linking participants to administrative data routinely collected in public health settings in Canada, we can follow up and infer the occurrence of morbidities such as cancers. These types of cohort resources allow for interrogation of pre-diagnosis biologic samples, as we described here using developments in cfDNA methylation profiling assays, and can be similarly extended to alternative emerging methodologies interrogating blood biomarkers such as cell-free RNA, proteins, and metabolites.

Several recent studies have profiled plasma cfDNA methylation profiles of breast cancers for early cancer detection, however these studies primarily sample from patients after clinical detection^{10–12}. To date, only one study has profiled breast cancer plasma collected prior to clinical detection, in that case using a single methylome marker, and reporting sensitivities between 5-12% and 88% specificity among samples collected two to three years before diagnosis²⁹. Comparatively, our predictive models achieve a mean sensitivity of 42.5% at 99% specificity for classifying breast cancer cases diagnosed three or more years after blood plasma profiling. Existing methylome profiling of plasma samples from post-diagnosis, and presumably more advanced breast cancer patients have typically noted better classification performance among HR negative breast cancers relative to HR positive²⁸. Owing to the low incidence rate of HR negative breast cancer in the OHS, we suspect a poorer predictive performance among HR negative breast cancers due to biasing toward selected features associated with HR positive breast cancers. Alternatively, considering that HR positive breast cancers typically have slower doubling times, less aggressive cancers may be present for longer but remain undetected by mammograms until reaching visible sizes allowing for a longer window of opportunity for detection at an early stage and age. Whereas aggressive cancers which develop and expand more rapidly, may have a shorter window of opportunity for detection at an early stage. Further, the batching of case and control groups are often not reported across early cancer detection studies and as such internal model performance can often be inflated if cases and controls are processed in separate batches. When case and control groups are perfectly confounded between batches, signals associated with technical artifacts can often drive separation of case and control groups in both training and testing samples, consequently conflating predictive performances³⁰. Accordingly, we profiled our cases with control samples between sequencing runs in this study, in addition to using a repeated cross-validation approach to estimate the uncertainty of predictive performances. However, false-positive predictions in our cohort may represent misclassifications of control samples with undetected underlying cancers, owing to variable follow-up duration among cancer-free controls (Extended Data Fig. 1).

Additionally, the following limitations of the current study should be considered when interpreting our findings. Firstly, 1.6 mL of plasma was used per participant for this study, which is a substantial amount of biobanked material, but larger plasma volumes would likely increase the number of ctDNA fragments captured and further improve detection sensitivity. Owing to the prospective nature of the OHS cohort, our sample sizes of pre-diagnosis cancers were limited by the incidence rate of the cancer among the study population with a cryopreserved blood sample, acknowledging that these incident cases will accrue with time. Additionally, not all cancer-free control samples were followed up for the same duration owing to our matching of controls to cases by sample collection time. While we followed all controls up to 2019 to ensure that they did not pass away due to other conditions and have had no history of cancer, it is possible that controls with shorter follow up times may have underlying undetected cancers that are yet to be diagnosed. Consequently, this may inflate the false positive rate by mislabelling control samples with undiagnosed cancers, which may also inflate false negative rate by reducing the power for detecting cancer specific DMRs if controls with undiagnosed cancers harbored the same hypermethylated regions with pre-diagnosis cases. Finally, existing studies profiling post-diagnosis breast cancer cfDNA methylomes have reported better predictive performance in HR- compared to HR+ breast cancers, contrary to our current findings. The pre-diagnosis breast cancer signatures we identified in this work could potentially be biased towards HR+ breast cancers, the majority subtype in our cohort and among breast cancers. Larger sample sizes will likely enable detection and comparisons of subtype-specific markers in the pre-diagnostic context or additional prognostic signatures.

Currently, breast cancer is one of few cancer types with a population screening tool owing to its associated reduction in mortality. Consequently, most breast cancers are typically diagnosed at stage I or II as seen in the OHS cohort and across the population. While mammograms are currently the gold standard for early breast cancer screening achieving a sensitivity and specificity of 92% in Ontario³¹, respectively, low patient compliance for subsequent screening, and low-dose radiation exposure may also increase the risk of future breast cancer development³². A liquid biopsy-based approach could not only enable simultaneous detection of multiple cancers, but also mitigate risks associated with radiographic imaging approaches, particularly for individuals preceding mammogram eligible ages. While it is unclear whether diagnoses prior to mammogram detection will further improve prognostic outcomes, detection of breast cancer signatures even at seven years prior to a stage I or II diagnosis presents promising results for early pre-symptomatic detection among other cancer types with no reliable screening tool. Additionally, interrogating genome-wide cfDNA methylation patterns can enable simultaneous interrogation of multiple cancer types and potentially other non-cancer conditions, whereas mammograms may be limited to cancer detection only in breast tissue. Indeed, future applications of liquid biopsies for early cancer detection will require identifying the tissue of origin of underlying cancers. Future investigations in the pre-diagnosis space with additional cancers from other tissues will allow for identifying tissue-specific markers and the development of tissue of origin classifiers, similar to those demonstrated in existing studies classifying post-diagnosis samples.

Methods

Patient Selection and plasma samples

Patient plasma samples were obtained from the Ontario Health Study (OHS) with protocols approved by the University of Toronto Health Sciences research ethics board (protocol #34088). Peripheral blood was drawn from OHS participants upon recruitment to the study, and 1.6 mL plasma was separated and collected within 48 hours, and immediately cryopreserved at the OHS Biobank. Participants in OHS that had developed breast cancer following recruitment to the study were identified by linking individuals who had provided a blood sample to the Ontario Cancer Registry through Cancer Care Ontario (CCO). Variables that were used for linkages included health insurance number, age, sex and name. At the time of linkages, cancer registry data had been made available through the Ontario Cancer Registry up until December 2017. Breast cancers were confirmed by histological analyses of tissue biopsies at the time of diagnosis, and immuno-histochemical tests for hormone receptor status were reported in the pathology records of breast cancer cases. A total of 167 OHS participants that donated biologics developed breast cancer following study recruitment. Blood plasma from 110 breast cancer participants was available and pulled from the biobank. Additionally, 1.6 mL plasma from 108 cancer-free controls matched to cases by age, sex, date of biologic collection, ethnicity, smoking status, and alcohol consumption frequency were also selected. Control participants that have not had a history of cancer prior to or following study enrolment and did not pass away from other comorbidities up to December 2017 were retained.

Next-generation Sequencing Library Construction and cfMeDIP-Seq protocol

The cfDNA was extracted from plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen). 5-10ng of cfDNA was used as input to generate methylated cfDNA libraries (IP libraries) along with an input control library (IC libraries). Quality of incoming cfDNA was assessed using the Fragment Analyzer (Agilent) following the manufacturers guidelines. 0.1ng of *Arabidopsis thaliana* DNA was added to samples prior to library preparation. Combined samples were prepared using the KAPA Hyper Prep library protocol (Roche), with standard End Repair & A-tailing and ligation of xGen Duplex Seq Adapter (IDT), followed by incubation at 4°C overnight. Unmethylated lambda (λ) DNA was added to partially completed IP libraries and enriched for methylated DNA using the MagMeDip Kit (Diagenode) and purified with the IPure Kit v2 (Diagenode). Sample indices were added to IP and IC libraries via PCR. Completed libraries were quantified by Qubit (Life Technologies) and Fragment Analyzer (Agilent). Both IP and IC libraries underwent shallow sequencing (~20,000 reads) on the MiSeq platform as a quality control step. IP libraries were sequenced to approximately 60M read pairs in 2x50bp mode on Novaseq platform (Illumina). All breast cancer samples were batched together with controls to mitigate batch effects between sequencing runs.

Raw Sequencing File Processing

Following sequencing, the FASTQ raw reads were adapter trimmed, with unique molecular identifiers (UMIs) appended to fastq headers using *UMI Tools* (version 0.3.3). The reads were then aligned to hg19 using *Bowtie2*³³ (version 2.3.5.1) in paired end mode at default settings. Aligned SAM files were

converted to BAM file format, indexed, and sorted using *SAM tools* (version 1.9)³⁴. Aligned reads were subsequently deduplicated according to alignment positions and UMIs using *UMI Tools*.

Quality Control and Sample Inclusion

One control sample was excluded from our study owing to mortality from non-cancer related causes during study follow up. Six controls were excluded due to diagnoses of cancer pre-disposing conditions that were identified from study follow up questionnaires during follow up. Three control samples were excluded owing to diagnosis of another cancer following sample collection and processing. Following library preparation, four samples were removed as no reads were generated during the MiSeq quality control step. We retained and analysed all samples with more than 10 million deduplicated reads. 39 samples were removed owing to Novaseq sequencing instrument failure that resulted in poor sequencing yields. To assess enrichment efficiency, the number of methylated and unmethylated Arabidopsis spike-ins aligned to F19K16 and F24B22 respectively were counted, and the proportion of methylated spike-ins generated out of the total spike-ins were calculated. Seven samples with less than 95% of spike-in reads that were methylated were excluded. An additional eight were samples owing to poor CpG enrichment assessed through GoGe (< 1.75) and relH enrichment scores (< 2.7) calculated using *MEDIPS* (R package version 1.12.0) were also removed³⁶. See Extended Data Table 1 for quality control metrics and sample information among remaining samples.

Computing cfMeDIP-Seq methylation signals

To identify regions with differential methylation between pre-diagnosis breast cancers and control cfDNA, coverage profiles were generated for each sample across 300 bp non-overlapping binned tiled windows using *MEDIPS*. To reduce background signal from non-tumour-derived cfDNA and reduce the feature search space, we leveraged publicly accessible data to filter for potentially informative regions. Regions frequently methylated in haematopoietic cells were inferred using whole genome bisulfite sequencing data of peripheral blood leukocytes ($n = 78$) from the International Human Epigenetics Consortium (IHEC)³⁶. We averaged the level of methylation across all CpG sites within the same 300 bp non-overlapping tiled window for each sample to infer the level of methylation within a specified region. Regions with an average of methylation level greater than 0.4 across all PBL samples were excluded. Remaining 300-bp bins with at least six or more CpG sites located at CpG islands, shores and shelves, or in FANTOM5 annotated promoters and enhancers were tested for differential methylation³⁷.

Repeated cross validation for differential methylation calling and predictive modelling

A 10-fold cross validation (CV) approach was used to evaluate the discriminatory performance of a methylation biomarker. First, the pre-cancer cases and control samples were divided into 10 approximately equal sized sets using stratified sampling, balancing the proportion of pre-diagnosis cases by years prior to diagnosis following blood collection in each fold set. Iteratively, for each fold in the CV procedure, one set was selected as the test set and the remaining nine sets were designated as the train set (comprising 10% and 90% of participants respectively). Within the train set, differential methylation

calling was performed using a Wald test of the regression coefficient from a negative binomial regression of cfMeDIP-Seq methylation level on train set case and control status using *DESeq2* (R package version 1.30.1), adjusting for batch and age³⁸. Additionally, we filtered out regions lowly methylated in cancer-free participants in the train set by identifying regions with a mean count across train set controls less than the mean count across all regions among train set samples. The remaining features with a $p < 0.05$ were retained and considered significant within each subsampling iteration. Across the 2000 (200 replicates of 10-fold CV) subsampled DMR calls, regions that passed all filter thresholds and were significant across at least 200 DMR calls ($> 10\%$; 485 regions) were retained and investigated.

Within each subsampling iteration, the top 200 hypermethylated features were used to construct a random forest model with *Caret* (R package 6.0)³⁹ from methylation counts of train set samples normalised by library size using *DESeq2*. The model performance was then assessed by applying the predictive model to the held-out test to obtain risk scores that reflect the proportion of decision trees that classify the sample as breast cancer. The 10-fold CV procedure was repeated 200 times to estimate the uncertainty in the analysis results. To assess the performance across the 200 repeats, we averaged the AUC and other performance metrics across the 10-folds for each CV repeat and calculated overall average performance and confidence intervals across the 200 repeated procedures.

TCGA Breast Cancer and Pan-cancer DNA Methylation Array

To identify whether cfDNA pre-diagnostic DMRs overlapped with DMRs between breast cancer and other tissue types, we conducted differential methylation analysis on TCGA 450k methylation array data from 976 paired healthy and normal tissue biopsy spanning 12 cancer types, and between publicly accessible PBL 450k methylation array data and TCGA breast cancer data⁴⁰. Solid cancer and normal tissue raw IDAT files were downloaded from the TCGA data portal, and PBL from the GeoExpression Omnibus (GSE87571 and GSE42861). IDAT files were processed to generate beta methylation values from IDAT files using *Minfi* (1.36.0 R package)⁴¹, and normalised using the *preprocessFunnorm* function. To test for differentially methylated CpG sites between paired healthy and tumour biopsies, an F-test was performed using the *DMPFinder* function from *Minfi* across 485,512 CpG sites. To avoid imbalances in DMR calling towards specific cancer types with more samples, a resampling without replacement of 5 paired healthy and tumour biopsies from each of the 12 cancer types was repeated 1000 times, and a median p-value across the 1000 repeats was calculated per probe site. An FDR correction was applied to median p-values, and CpG sites with a median FDR q-value below 0.01 and median absolute difference in methylation of greater than 0.1 were retained as candidate markers for discriminating between tumour and healthy tissue. Differential methylation calling between all TCGA breast cancer and TCGA breast normal tissue, as well as between breast cancer tissue and PBL was also conducted using the *DMPFinder* function between all samples from each respective group to identify additional breast cancer specific markers. To infer whether cfDNA markers were reflective of breast cancer tissue methylomes, we identified 112 out of 207 significant DMRs among pre-diagnostic cfDNA markers overlapping with 247 CpG sites on 450k methylation array probes to cluster breast cancer, breast normal, PBL and TCGA tissue samples using tSNE.

Transcription factor and gene set enrichment

Transcription factor binding sites among significant DMRs were located using *LOLA* (R package version 1.20.0)⁴², with ChIP-Seq data from *Encode* used as a reference for known transcription factor binding sites⁴³. Permutation testing was performed by randomly subsampling 485 regions 200 times across the genome to obtain a distribution of expected number of overlaps with TFBS of interest. The number of overlapping binding sites among randomly subsampled regions and among significant DMRs for each transcription factor of interest were z-score normalised. Gene set enrichment analysis among the 485 significantly hypermethylated regions was performed using the R package *rGREAT* (R package version 1.22.0) with default settings to identify enriched gene ontology molecular functions⁴⁴. Gene sets with a binomial test and FDR adjusted p-value < 0.1 were considered significant.

Time-dependent model assessment

Owing to the outcome and age dependent sampling used in the study, the artificial case-to-non-case ratio in our sample was not representative of the Canadian adult population. We calculated sampling weights to adjust for this sampling bias in the time-dependent model assessment analysis using age specific cumulative breast cancer incidence rates from the Canadian Cancer Registry, in addition to all-cause mortality rates in Ontario reported by Statistics Canada. Often, case-non-case samples are analyzed without regard for the variable length of follow-up (time-to-event for cases, and time-to-censoring or loss to follow-up for non-cases). In cohort study samples, event times of cases may be longer than non-case follow-up times, and classification accuracy is assessed for intermittent follow-up times, both of which require time-dependent AUCs to summarize the discriminatory capacity of a model or marker. The predictive performance was assessed using time-dependent receiver operating characteristic (ROC(t)) curves, corresponding area under the curve (AUC(t)) estimates, accounting for the sampling weights. Cross-validated AUC estimates were obtained by taking the average of the AUC estimates from all “left-out” folds within a 10-fold CV replicate. The mean and the 2.5% and 97.5% quantiles were obtained from the AUC estimates across the set of 200 replicates of the CV procedure. As a summary measure of the classification accuracy across all follow-up times observed in the sample, we calculated a weighted concordance index; defined as the probability that, for any two randomly chosen participants, the observation with the shorter time to diagnosis also had the larger risk score assigned by the random forest classifiers. Pairs within the calculation were weighted to adjust for the study sampling⁴⁵. Concordance, like the area under the curve (AUC) statistic, measures a model’s ability to discriminate between cases and non-cases, but does not address absolute probabilities assigned to each class.

Declarations

Data Availability

The data generated by the present study are available from OHS (<https://www.ontariohealthstudy.ca/researchers/data-access-process/>) or the CanPath portal (<http://portal.partnershipfortomorrow.ca>) upon

request. TCGA 450k methylation array and RNA-seq data for cancer and matched normal tissues are publicly available through the Genomics Data Commons portal (GDC, <https://portal.gdc.cancer.gov/>) and GDC legacy archive (<https://portal.gdc.cancer.gov/legacy-archive/search/f>). 450k methylation array for peripheral blood leukocytes are available through the Gene Expression Omnibus (accession numbers GSE87571 and GSE42861). Whole genome bisulphite data of peripheral blood leukocytes are publicly available through the international human epigenetics consortium (<https://epigenomesportal.ca/ihec/>).

Code Availability

The codes used to process cfMedIP-Seq data and perform analyses are available at <https://github.com/nickcheng96/Cheng-et-al.-Pre-diagnosis-BRCA-cfMeDIP-Seq>.

Acknowledgments

We would like to acknowledge the Genomics Research Platform team at OICR for performing the cfMeDIP-Seq assay on the plasma samples, as well as the insightful comments on the study from members of the Ontario Institute for Cancer Research. Biological materials were stored at the Ontario Health Study Biobank, which is supported by the Ontario Institute for Cancer Research through funding provided by the Government of Ontario and a Genome Canada grant (OGI-136) samples were excluded from analysis owing to PA. Parts of this material are based on data and information provided by Ontario Health, and includes data received by Ontario Health from the Canadian Institute for Health Information (CIHI) and the Ministry of Health (MOH). The opinions, reviews, views and conclusions reported in this publication are those of the authors and do not necessarily reflect those of Ontario Health, CIHI, and/or the MOH. No endorsement by Ontario Health, CIHI, and/or the MOH is intended or should be inferred.

Competing interest statement

D.D.C and S.B are listed as inventors/contributors on patents filed related to the cfMeDIP-seq technology. D.D.C. received research funds from Pfizer and Nektar therapeutics. D.D.C. and S.B. are co-founders and shareholders of Adela. All the other authors declare no competing interest.

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians* **69**, 7-34 (2019).
2. Smith, R. A. *et al.* Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians* **69**, 184-210 (2019).

3. Ebell, M. H., Thai, T. N. & Royalty, K. J. Cancer screening recommendations: an international comparison of high income countries. *Public Health Rev.* **39**, 7 (2018).
4. Lehman, C. D. *et al.* National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49-58 (2017).
5. Guertin, M. *et al.* Mammography Clinical Image Quality and the False Positive Rate in a Canadian Breast Cancer Screening Program. *Can Assoc Radiol J* **69**, 169-175 (2018).
6. Klarenbach, S. *et al.* Recommendations on screening for breast cancer in women aged 40-74 years who are not at increased risk for breast cancer. *CMAJ* **190**, E1441-E1451 (2018).
7. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926-930 (2018).
8. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579-583 (2018).
9. 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).
10. Zhang, X. *et al.* Circulating cell-free DNA-based methylation patterns for breast cancer diagnosis. *npj Breast Cancer* **7**, 106 (2021).
11. Cristall, K. *et al.* A DNA methylation-based liquid biopsy for triple-negative breast cancer. *npj Precision Oncology* **5**, 53 (2021).
12. Moss, J. *et al.* Circulating breast-derived DNA allows universal detection and monitoring of localized breast cancer. *Annals of Oncology* **31**, 395-403 (2020).
13. Ontario Health Study. (2021); Available from: <https://www.ontariohealthstudy.ca/>
14. Shen, S. Y., Burgener, J. M., Bratman, S. V. & Carvalho, D. D. D. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nature Protocols* **14**, 2749-2780 (2019).
15. Mohn, F., Weber, M., Schübeler, D. & Roloff, T. Methylated DNA immunoprecipitation (MeDIP). *Methods Mol Biol* **507**, 55-64 (2009).
16. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications* **9**, 1-12 (2018).
17. Beleites, C. *et al.* Variance reduction in estimating classification error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems* **79**, 91-100 (2005).
18. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* **19**, 371-384 (2018).
19. Das, T. K., Sangodkar, J., Negre, N., Narla, G. & Cagan, R. L. Sin3a acts through a multi-gene module to regulate invasion in *Drosophila* and human tumors. *Oncogene* **32**, 3184-3197 (2013).
20. Chen, L. Tumor suppressor function of RUNX3 in breast cancer. *J. Cell. Biochem.* **113**, 1470-1477 (2012).
21. Damaschke, N. A. *et al.* CTCF loss mediates unique DNA hypermethylation landscapes in human cancers. *Clinical Epigenetics* **12**, 80 (2020).

22. Werner, S. et al. Iroquois homeobox 2 suppresses cellular motility and chemokine expression in breast cancer cells. *BMC Cancer* **15** (2015).
23. Hsu, N. C. et al. Methylation of BRCA1 promoter region is associated with unfavorable prognosis in women with early-stage breast cancer. *PloS one* **8**, e56256 (2013).
24. Fernández-Cabezudo, M. J. *et al.* Deficiency of mitochondrial modulator MCJ promotes chemoresistance in breast cancer. *JCI Insight* **1** (2016).
25. Schaafsma, E., Fugle, C. M., Wang, X. & Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. *Br. J. Cancer* **125**, 422-432 (2021).
26. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337-344 (2000).
27. Acheampong, T., Kehm, R. D., Terry, M. B., Argov, E. L. & Tehranifar, P. Incidence Trends of Breast Cancer Molecular Subtypes by Age and Race/Ethnicity in the US From 2010 to 2016. *JAMA Netw Open* **3**, e2013226 (2020).
28. Liu, M. C. *et al.* Breast cancer cell-free DNA (cfDNA) profiles reflect underlying tumor biology: The Circulating Cell-Free Genome Atlas (CCGA) study. *JCO* **36**, 536 (2018).
29. Widschwendter, M. *et al.* Methylation patterns in serum DNA for early identification of disseminated breast cancer. *Genome Medicine* **9**, 115 (2017).
30. Soneson, C., Gerster, S. & Delorenzi, M. Batch Effect Confounding Leads to Strong Bias in Performance Estimates Obtained by Cross-Validation. *PloS one* **9**, e100335 (2014).
31. Ontario Health. The Ontario Cancer Screening Performance Report 2020. *Cancer Care Ontario*. (2020).
32. Boice, J. D., Harvey, E. B., Blettner, M., Stovall, M. & Flannery, J. T. Cancer in the Contralateral Breast after Radiotherapy for Breast Cancer. *New England Journal of Medicine* **326**, 781-785 (1992).
33. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
35. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics (Oxford, England)* **30**, 284-286 (2014).
36. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145-1149 (2016).
37. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Scientific data* **4**, 170112 (2017).
38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

39. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1-26 (2008).

40. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113-1120 (2013).

41. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).

42. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589 (2016).

43. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831 (2012).

44. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

45. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543-2546 (1982).

Tables

Table 1 is in the supplementary files section.

Figures

Figure 1

Overview of study design and analytical approach. (a) Outline of participant recruitment and blood plasma sample selection process in the Ontario Health Study (OHS). Blood plasma of 218 participants in the OHS was pulled from the Ontario biobank, of which 150 passed all quality control metrics and inclusion criteria (b) Timeline of blood plasma collection, breast cancer diagnosis and last mammogram prior to biologic collection across incident breast cancer cases in OHS. Each row represents an individual

breast cancer case (n = 78). Brown dots indicate the time of biologic sample, blue dots represent the last self-reported mammogram prior to blood collection, and red dots indicate time of breast cancer diagnosis

(c) Count of breast cancer cases at each incremental timepoint between sample collection and breast cancer diagnosis. Colors indicate the stage at diagnosis across pre-diagnosis breast cancer cases.

Figure 2

Schematic of the analytical approach performed to classify pre-diagnosis breast cancer samples using differentially methylated regions in cfDNA. Pre-diagnosis breast cancer cases and cancer-free controls were partitioned into 10-fold splits, balancing the number of cases by years prior to diagnosis evenly among each fold. Iteratively, nine-folds (blue and purple boxes) are selected as train set samples for identifying significantly ($p < 0.05$) differentially methylated regions in cfDNA, and to build a binary random forest classifier for identifying individuals with undetected breast cancers using the top 200 most significant regions hypermethylated in pre-diagnosis cases. The one remaining held-out test fold (yellow box) that was not involved in any aspect of feature selection or model building was used to evaluate the classifier performance. This process was iteratively repeated 10 times, such that each fold was the held-out test fold once. We repeated this 10-fold CV split strategy 200 times, each time with random sample partitioning into folds.

Figure 3

Significantly hypermethylated regions between pre-diagnosis breast cancer and cancer-free control cases.

(a) Principal component analysis of the top 50 hypermethylated regions between pre-diagnosis breast cancer cases versus controls. Colors indicate time between sample collection and diagnosis. **(b)** Enrichment of CpG dense regions differentially methylated in breast cancer tissue among 485 significantly hypermethylated cfDNA DMRs. Enrichment measured using a permutation analysis of the observed overlap between significantly hypermethylated within genomic regions of interest relative to a distribution of the expected overlap generated by random subsampling of the genome. Counts indicate the number of significant DMRs overlapping with respective genomic regions differentially methylated in TCGA breast cancer tissue. **(c)** Enrichment for transcription factor binding sites among cfDNA hypermethylated regions in pre-diagnostic breast cancer cases measured using a permutation analysis. Counts indicate the number of significant DMRs that are located in binding sites of respective transcription factors. **(d)** Absolute change in promoter methylation and log2 fold-change in gene expression between TCGA tumour and normal breast cancer tissue among regions identified from pre-diagnosis breast cancer cfDNA. Each color represents a PAM50 breast cancer subtype or breast normal tissue **(e)** False discovery rate adjusted enriched gene ontology molecular functions among significantly hypermethylated cfDNA DMRs.

Figure 4

Predictive performance of pre-diagnosis breast cancer cfDNA DMRs for classifying individuals (n = 78 cases, n = 72 controls) with early breast cancer (a) Mean receiver operator characteristics for predicting pre-diagnosis breast cancer cases across 200 10-fold cross validated repeats. Shading indicates 95% confidence intervals (b) Mean time-dependent area under receiver operating curves (AUC(t)) that takes into consideration the time to diagnosis of breast cancer cases and control follow-up times to quantify the discriminative ability of cfDNA methylation markers at various time points prior to diagnosis. AUC(t) at each timepoint is weighted for the true cumulative age-specific breast cancer incidence rates in Canada. The concordance index is the proportion of concordant pairs of samples in which the sample with the higher risk score develops or is diagnosed with breast cancer sooner compared to samples with a lower risk score. (c) Mean receiver operator characteristics and (d) AUC(t) for predicting pre-diagnosis breast cancer cases diagnosed at stage I. (e) Mean receiver operator characteristics and (f) AUC(t) for predicting pre-diagnosis hormone receptor positive breast cancer cases. (g) Mean receiver operator characteristics and (h) AUC(t) across pre-diagnosis breast cancer case subgroups. Bands indicate 95% confidence intervals.

Figure 5

Predictive performance of pre-diagnosis breast cancer cfDNA DMRs for classifying individuals with a negative mammogram screen within one year of blood plasma collection (a) Mean receiver operator characteristics and (b) AUC(t) across 200 10-fold cross-validated repeats. Shading indicates 95% confidence intervals (c) Timepoints of last mammogram screen before biologic collection (blue dots), blood plasma collection (brown dots) and breast cancer diagnosis (red dots). Each row represents are breast cancer case.

Supplementary Files

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

- [ExtendedData.docx](#)
- [Table1.docx](#)