

RESEARCH

Open Access



# Buffy coat signatures of breast cancer risk in a prospective cohort study

Felicia Fei-Lei Chung<sup>1,2\*\*†</sup>, Sandra González Maldonado<sup>3†</sup>, Amelie Nenc<sup>4†</sup>, Liacine Bouaoun<sup>1†</sup>, Vincent Cahais<sup>1</sup>, Cyrille Cuenin<sup>1</sup>, Aurelie Salle<sup>1</sup>, Theron Johnson<sup>3</sup>, Bekir Ergüner<sup>4</sup>, Marina Laplana<sup>5,6</sup>, Paul Datlinger<sup>4</sup>, Jana Jeschke<sup>7</sup>, Elisabete Weiderpass<sup>1</sup>, Vessela Kristensen<sup>8</sup>, Suzette Delalogue<sup>9</sup>, François Fuks<sup>7</sup>, Angela Risch<sup>5,10,11</sup>, Akram Ghantous<sup>1</sup>, Christoph Plass<sup>5</sup>, Christoph Bock<sup>4,12†</sup>, Rudolf Kaaks<sup>3†</sup> and Zdenko Herceg<sup>1\*†</sup>

## Abstract

**Background** Epigenetic alterations are a near-universal feature of human malignancy and have been detected in malignant cells as well as in easily accessible specimens such as blood and urine. These findings offer promising applications in cancer detection, subtyping, and treatment monitoring. However, much of the current evidence is based on findings in retrospective studies and may reflect epigenetic patterns that have already been influenced by the onset of the disease.

**Methods** Studying breast cancer, we established genome-scale DNA methylation profiles of prospectively collected buffy coat samples ( $n = 702$ ) from a case–control study nested within the EPIC–Heidelberg cohort using reduced representation bisulphite sequencing (RRBS).

**Results** We observed cancer-specific DNA methylation events in buffy coat samples. Increased DNA methylation in genomic regions associated with SURF6 and REXO1/CTB31O20.3 was linked to the length of time to diagnosis in the prospectively collected buffy coat DNA from individuals who subsequently developed breast cancer. Using machine learning methods, we piloted a DNA methylation-based classifier that predicted case–control status in a held-out validation set with 76.5% accuracy, in some cases up to 15 years before clinical diagnosis of the disease.

**Conclusions** Taken together, our findings suggest a model of gradual accumulation of cancer-associated DNA methylation patterns in peripheral blood, which may be detected long before clinical manifestation of cancer. Such changes may provide useful markers for risk stratification and, ultimately, personalized cancer prevention.

<sup>†</sup>Felicia Fei-Lei Chung, Sandra González Maldonado, Amelie Nenc, and Liacine Bouaoun are Joint first authors.

<sup>†</sup>Christoph Bock, Rudolf Kaaks, and Zdenko Herceg are Joint last authors.

\*Correspondence:

Felicia Fei-Lei Chung  
feliciacfl@sunway.edu.my  
Zdenko Herceg  
hercegz@iarc.who.int

<sup>1</sup> International Agency for Research On Cancer (IARC), 25 avenue Tony Garnier, CS 90627, 69366 Lyon, France

<sup>2</sup> Department of Medical Sciences, School of Medical and Life Sciences, Sunway University, 5, Jalan Universiti, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia

<sup>3</sup> Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

<sup>4</sup> CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

<sup>5</sup> Division of Cancer Epigenomics, German Cancer Research Center, Heidelberg, Germany

<sup>6</sup> Department of Basic Medical Sciences, University of Lleida, IRBLleida, 25198 Lleida, Spain

<sup>7</sup> Laboratory of Cancer Epigenetics, Université Libre de Bruxelles (ULB), Brussels, Belgium

<sup>8</sup> Faculty of Medicine, Institute for Clinical Epidemiology and Molecular Biology, University of Oslo, Oslo, Norway

<sup>9</sup> Department of Cancer Medicine, Institut Gustave Roussy, Villejuif, France

<sup>10</sup> Department of Biosciences and Medical Biology, Allergy-Cancer-BioNano Research Centre, University of Salzburg, 5020 Salzburg, Austria

<sup>11</sup> Cancer Cluster Salzburg, Salzburg, Austria

<sup>12</sup> Medical University of Vienna, Institute of Artificial Intelligence, Center for Medical Data Science, Vienna, Austria



**Keywords** Epigenetics, DNA methylation, Cancer risk markers, Breast cancer, Prospective cohort

## Background

Cancer is a leading cause of death worldwide and has been described as the single most important barrier to increasing life expectancy in the twenty-first century [1]. While the development of effective screening procedures has allowed for early detection of malignant lesions and reductions in cancer-related mortality [2], few early detection tests have been effective in reducing cancer-specific morbidity to date [3]. There is a need to re-examine the limitations of the current “one-size-fits-all” approach to cancer screening and to move towards more personalized approaches for prevention and early detection [4].

One strategy towards addressing this challenge is to integrate molecular markers in the generation of risk stratification profiles [2, 4]. Epigenetic markers have been put forward as important indicators of cancer risk, and they are highly attractive options in clinical practice because of their technical stability [4, 5]. Epigenetic measures of biological age, in particular, have been associated with cancer-related mortality [6–11] and have great potential utility as early biomarkers of disease risk [12]. Multiple studies to date have established that alterations in DNA methylation can be detected in DNA isolated from the peripheral blood of patients with cancer [13–17]. Recent reports combining epigenomic analyses with machine learning classifiers were able to infer not only the presence of tumours but also the tissue of origin or subtype of the tumours [18–23]. Although these findings offer promising evidence for the utility of epigenetic events as biomarkers or predictors of cancer, these studies are retrospective in nature, reporting on methylation markers that are detectable upon or after diagnosis.

To add value as an early detection or risk stratification strategy, proposed assays should be non-invasive and capable of detecting cellular alterations before the disease progresses to the lower detection limit of conventional screening modalities. To date, reports indicate that epigenetic markers can be detected in prospectively collected from apparently healthy individuals that are later diagnosed with breast [24, 25] and ovarian cancers [26], suggesting that the DNA methylation profile in peripheral blood may be altered years before the tumour is clinically detected. A report from the Taizhou Longitudinal Study revealed that an epigenomics-based blood test could identify stomach, oesophageal, colorectal, lung, or liver cancer in apparently healthy individuals up to 4 years before diagnosis [27]. However, separate meta-analyses on similar pre-diagnostic samples reported no

associations between risk of breast cancer [28] or gastric cancer [29] and DNA methylation measured at individual CpG sites. More research is warranted to better understand the circumstances under which epigenomics-based tests could be best utilized.

In the present study, we established genome-scale DNA methylation profiles of buffy coat samples from a nested case–control prospective study using reduced representation bisulphite sequencing (RRBS) to identify differentially methylated regions (DMRs) in breast cancer cases compared with controls. We observed that a Prediction Analysis for Microarrays (PAM) classification algorithm could discriminate individuals who developed breast cancer from those who did not. The final PAM model was tested on a held-out validation set, in which it was able to predict the occurrence of cancer in individuals months to years before clinical diagnosis of the disease.

## Results

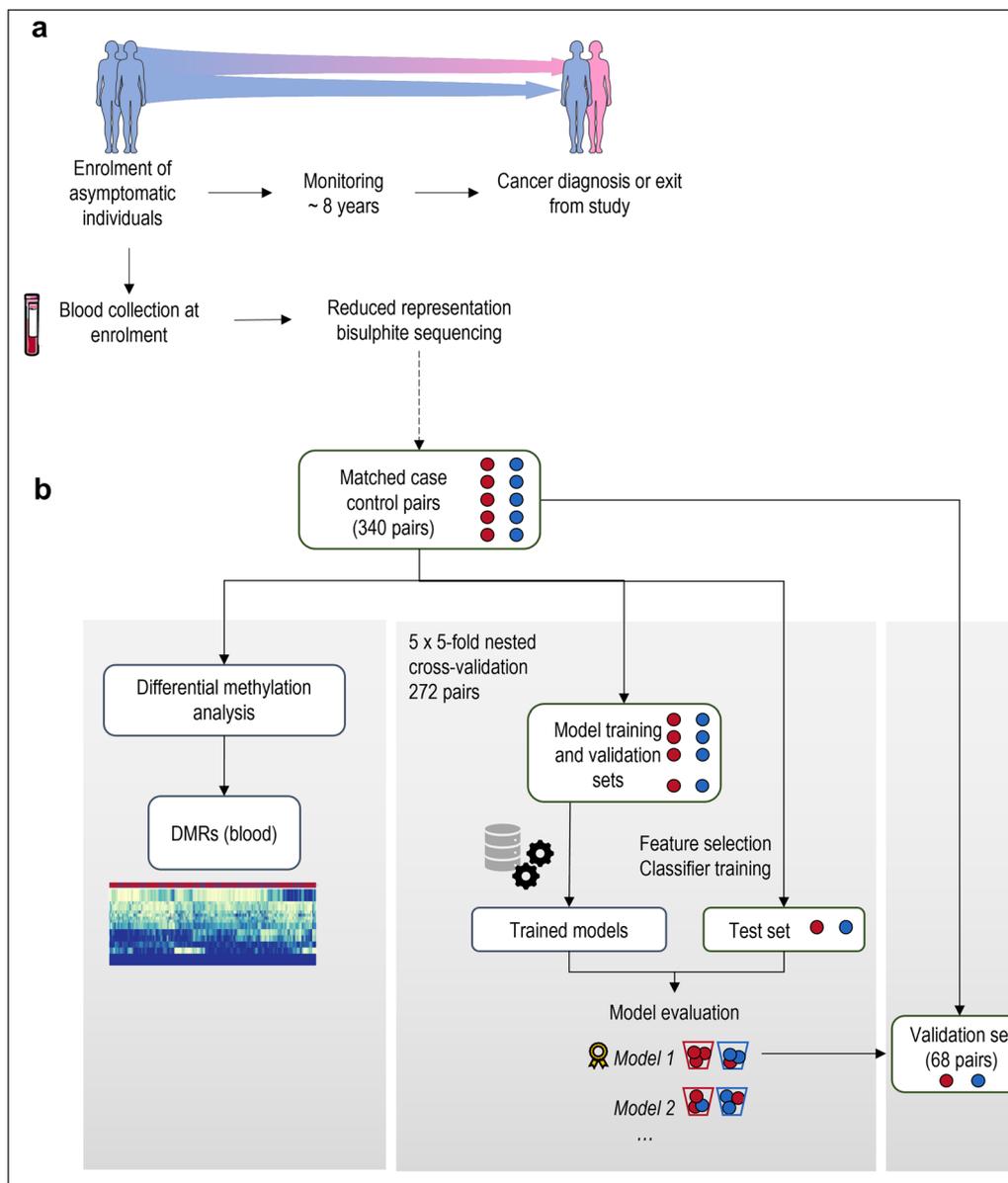
### Study design

Samples from the EPIC-Heidelberg cohort, a sub-cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC), study were used to construct a nested case–control study (study design is illustrated in Fig. 1). Blood samples were collected at enrolment from apparently healthy participants, from which buffy coat fractions were processed to yield a dataset of 702 RRBS profiles from 696 individuals. The final dataset consisted of 340 matched case–control pairs. Cohort characteristics are described in Additional file 1: Tables S1 and S2.

For predictive model development, 272 randomly selected matched pairs (80%) constituted a primary set that was used for model development and evaluation, and a set of 68 pairs (20%) was held out as a model validation set. Baseline cohort characteristics of the model development and validation sets are listed in Additional file 1: Table S2 and the distributions are graphed in Additional file 2: Fig. S1.

### Differentially methylated regions detected in prospectively collected buffy coat samples

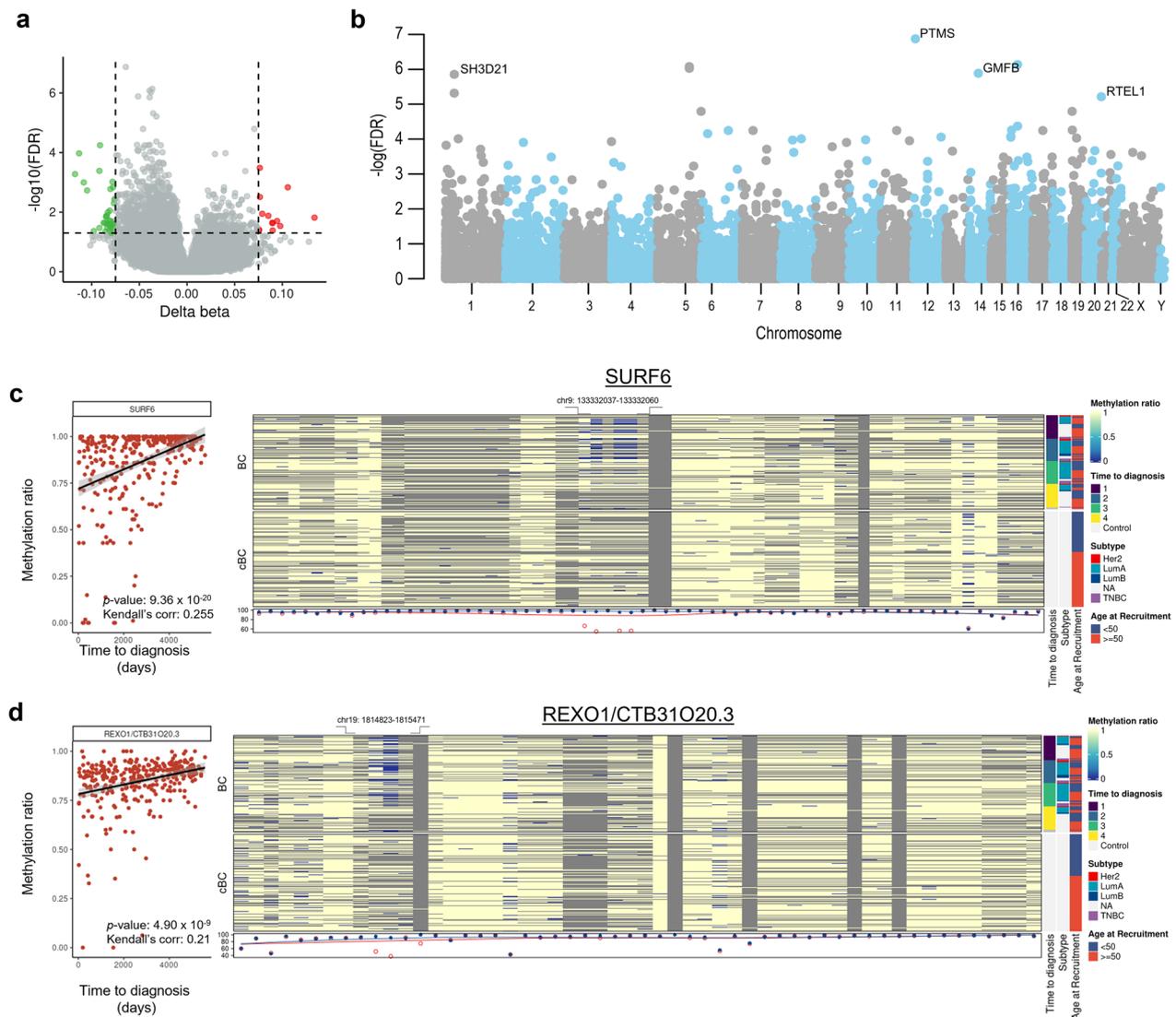
Paired differential analyses between cases and controls yielded 187 significantly differentially methylated genomic regions associated with 165 genes (false discovery rate [FDR]-adjusted  $p$  value < 0.05, absolute mean difference in beta values > 0.075). The full list of DMRs is given in Additional file 3: Table S3, and a representative volcano and Manhattan plot illustrating results of comparisons within gene promoter regions is shown in



**Fig. 1** Schematic of the study design and analytical methods. **a** A nested case–control study was constructed within the EPIC-Heidelberg cohort. Blood samples and lifestyle information were collected from apparently healthy participants upon enrolment. Participants who were diagnosed with breast cancer over the course of follow-up were matched on a one-to-one basis with individuals who were observed to be cancer-free over the study period. Buffy coat lysates derived from blood samples collected at enrolment were analysed by reduced representation bisulphite sequencing (RRBS). **b** Fivefold nested cross-validation was used to train and evaluate classifier models for their ability to discriminate individuals who developed breast cancer from those who were cancer-free over the follow-up period. The best-performing model was selected and trained on the full cross-validation dataset of 340 pairs, to finalize model parameters. The final model was used to predict case–control status in a held-out validation set of 68 matched pairs. Differential methylation analyses and functional enrichment analyses were conducted in parallel. This figure uses images from Servier Medical Art licensed under a Creative Commons Attribution 3.0 Unported License (<https://creativecommons.org/licenses/by/3.0/>)

Figs. 2a and 2b, respectively. When differential methylation analysis was conducted on pairs representing women diagnosed at above 50 years of age (representing post-menopausal breast cancer), 154 DMRs were identified,

corresponding to 128 known genes (Additional file 4: Table S4). Notably, 104 of these regions, corresponding to 65 known genes, overlapped with DMRs identified in the main analysis with all matched pairs.



**Fig. 2** Differentially methylated sites detected in prospectively collected blood samples from individuals who develop cancer within the study timeframe when compared with age- and lifestyle-matched control individuals. **a** Volcano plot illustrating results of differential methylation analyses comparing regions annotated to promoters between case and control buffy coat samples. **b** Manhattan plot illustrating the chromosomal locations of the top differentially methylated sites between cases and controls. Scatterplots demonstrating a positive correlation between DNA methylation levels at **c** SURF6 and **d** REXO1/CTB31O20.3 and length of time to diagnosis suggest that progressive demethylation at these sites could be indicative of early carcinogenesis. CpG sites that were found to be driving this relationship are highlighted in the regional heat map on the right

This included hypomethylation in genomic regions associated with oestrogen-related receptor beta (ESRRB) and the F-box protein member FBOX38 (Fig. 2a). Notably, ESRRB is a nuclear receptor and transcription factor which binds to the oestrogen-related receptor response element and is a key regulator in the reprogramming of pluripotent stem cells [30, 31] and glucocorticoid receptor signalling [32], whereas F-box proteins are members of the ubiquitin-protein E3 ligase family that play an important role in cell cycle regulation [33]. Pathway

enrichment analysis of the genes associated with the 187 DMRs identified from the case–control comparison indicated that there were no significantly enriched gene ontologies or pathways after correction for multiple testing (Additional file 5: Table S5). Similar analysis for overlapping genes of the main analysis (all case–control pairs) and the post-menopausal pairs revealed significant enrichment for the carbohydrate:proton symporter activity GO Molecular Function term (Additional file 6: Table S6). The hypermethylated and hypomethylated

regions of the main analysis (Additional file 3: Table S3) were significantly depleted for FANTOM5 enhancer regions identified in the GM12878 lymphoblastoid cell line relative to the total dataset (Fisher’s exact test,  $p < 0.05$ , Additional file 7: Fig. S2a). Similarly, the DMRs were depleted for promoter regions and were enriched for 1 to 5 kb regions and exonic regions relative to the total dataset (Additional file 7: Fig. S2b).

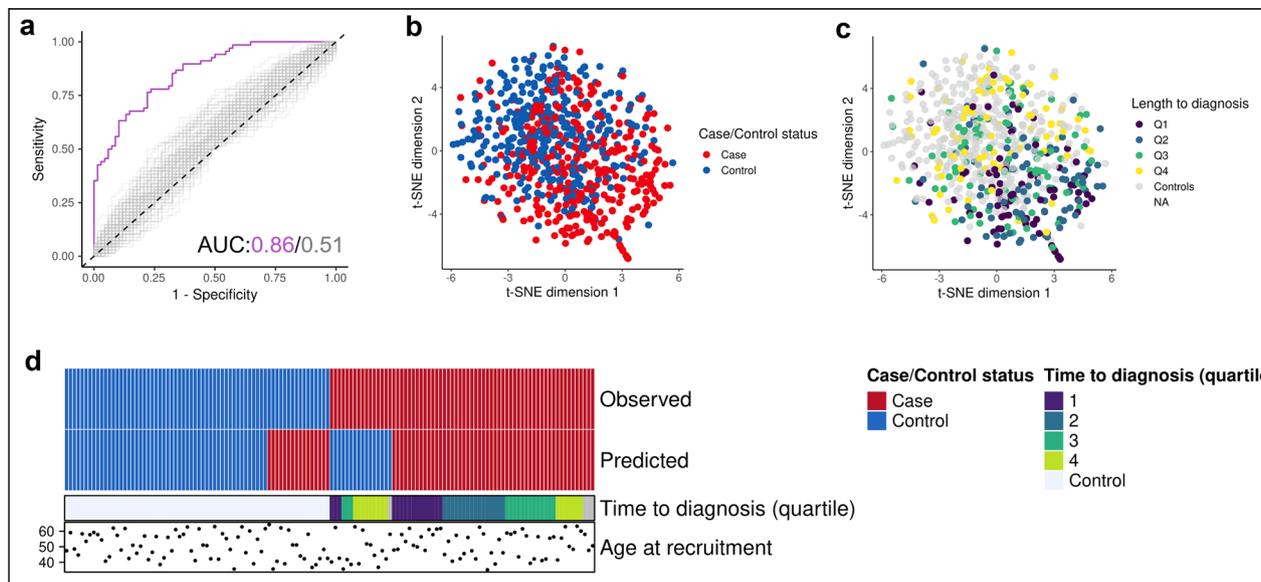
Of the 187 DMRs, 75 were significantly correlated with time to diagnosis in breast cancer cases (FDR-adjusted  $p$  value  $< 0.05$ , Additional file 3: Table S3). Regions most significantly correlated with time to diagnosis include SURF6 (Fig. 2c) and REXO1/CTB31O20.3 (Fig. 2d, Additional file 8: Table S7). CpG sites within the regions chr9: 133,332,037–133,332,060 for SURF6 and chr19: 1,814,823–1,815,471 for REXO1 were lowly methylated in cases that were diagnosed within 21–2665 days after recruitment (i.e. within the first and second quartile of patients by time to diagnosis), whereas higher levels of methylation were detected in matched controls and cases diagnosed more than 2665 days after recruitment.

**Identification of a panel of epigenetic predictors for breast cancer risk in RRBS dataset**

Several classifiers were tested for their ability to discriminate between cases and controls (a schematic of the approach is illustrated in Fig. 1b) using fivefold

cross-validation. The PAM classifier was the best-performing classifier overall when evaluated based on area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, and specificity (Additional file 1: Table S8, Additional file 10: Fig. S3). The PAM model used 49 genomic regions, corresponding to 38 known or predicted genes (Additional file 9: Table S9).

The PAM model was used to predict case–control status in the held-out set of 68 case–control pairs that were not used at any point during model development. The classifier correctly predicted case–control status in 52 of 68 cases, corresponding to an accuracy of 76.5%. The corresponding ROC curve and AUC statistic are shown in Fig. 3a, against a background of 100 label-shuffled datasets subjected both to the same feature selection (RFE) and classifier training process. The 49 predictive genomic regions used in the PAM classifier were used to generate a  $t$ -distributed stochastic neighbour embedding ( $t$ -SNE) plot, which showed considerable overlap between the case and control clusters (Fig. 3b). The cases most distinct from the controls were derived primarily from participants in the first and second quartiles by time to diagnosis (Fig. 3c). 13 of the 16 misclassified samples were in the third or fourth quartile of duration from sample collection to diagnosis, suggesting that the time to diagnosis could be an important factor influencing the performance of the predictors (Fig. 3d).



**Fig. 3** Prediction of case–control status in prospectively collected blood samples using a PAM classifier. **a** The receiver operating characteristic curve and the corresponding area under the curve (AUC) statistics for the PAM classifier applied on the validation cohort, against a background of 100 label-shuffled control datasets that were subjected to the same model training and testing process. A  $t$ -distributed stochastic neighbour embedding ( $t$ -SNE) plot was generated using the 49 genomic regions used in the PAM classifier, coloured by **b** case–control status and **c** length of time from sample collection to diagnosis (by quartile). **d** Schematic of the classification results from the final PAM model on the held-out validation set alongside length of time to diagnosis (quartiles)

## Discussion

Epigenetic mechanisms play an integral role in coordinating spatiotemporal gene expression, enabling the emergence of diverse cell type-specific phenotypes [20]. In cancers, one of the most well-described epigenetic aberrations is DNA hypomethylation within intergenic regions and/or partially methylated domains, punctuated by hypermethylation of CpG-dense regions [34–37].

Differentially methylated regions in the buffy coat samples include regions associated with surfet locus protein 6 (SURF6), deregulation of which have been reported in the peripheral blood cells of breast cancer patients [38]; ESRRB, a key regulator of stem cell pluripotency [30] and self-renewal [39, 40]; and FBXO38, which mediates the ubiquitination and degradation of the substrate programmed cell death protein 1 (PD-1) [41]. We observed that DNA methylation levels in a subset of these DMRs were significantly associated with length of time to diagnosis, lending confidence to our hypothesis that gradual alterations in DNA methylation states indicative of early phases of tumour development are detectable in the blood (buffy coat) samples prior to clinical diagnosis. Similar observations were reported by Xu et al. (2020), whereby 42.6% of the CpG sites found to be differentially methylated between cases and controls were significantly correlated with time to diagnosis [24]. Here, the authors opined that this progressive divergence suggests that the detected alterations to blood DNA methylation are an early response to tumour development, rather than a long-term marker of breast cancer susceptibility, where in the latter case blood DNA methylation alterations would be expected to be independent of time to diagnosis.

Additionally, we postulate that as these epigenetic alterations were detected in buffy coats, epigenetic alterations in these regions may not necessarily reflect the molecular/cellular alterations leading to or arising from carcinogenesis in the target tissue, and could instead be reflections of molecular/cellular processes associated with the early stages of tumour development, such as chronic inflammation and accelerated ageing, deleterious exposures, or any combination of the above, the identification of which is beyond the scope of the current study. Additionally, the observed epigenetic alterations may also result from changes to the composition of cell types in the buffy coats analysed, a factor which is documented to be sensitive to chronic and acute stressors [42, 43]. While methodologies for deconvoluting cell-type composition have been well-established for array-based datasets, similar methodologies have yet to be used widely with RRBS datasets. Due to the fact that methodologies for deconvolution are not suited to be applied across platforms [44], cell-type deconvolution was not conducted in this study.

Thus, the epigenetic alterations observed may have arisen due to alterations in cell-type composition, or due to large epigenetic alterations in specific cell types. Regardless, the processes resulting in the alterations observed in this study should be of a stable or persistent nature as the resulting epigenetic events persist through the continuous renewal of blood cells. Whether these alterations occur as a result of persistent “provoking conditions”, or stable epigenomic alterations in progenitor cells or in long-lived specialized blood cells (e.g. memory lymphocytes) [43] remains to be determined. Further studies are thus warranted to determine if the isolation of specific cell populations for DNA methylation analysis will prove beneficial for identifying robust biomarkers in prospectively collected blood samples [43].

While predictor performance reported in this study is lower than that in previous studies, we note that the current study uses prospectively collected blood samples, compared with the majority of current reports, in which analyses were conducted on samples from participants already diagnosed with the disease [20–22]. Moreover, the aforementioned studies utilized cell-free DNA isolated from plasma or serum as opposed to buffy coat samples.

This finding also follows two studies exploring the epigenetic differences between cases and controls in prospective studies, which reported contrasting results [24, 28]. We emphasize that these findings do not indicate that the use of circulating biomarkers could be applied in diagnosis of the disease, but could serve as an important component of personalized risk-based early prevention strategies [4]. Because it is accepted that breast cancer risk can be best predicted by a combination of parameters including age, genetic variants, mammographic breast density, reproductive history and lifestyle factors, the present study provides a novel epigenetic risk classifier and demonstrates the potential utility of DNA methylation markers in detecting early cellular alterations involved in tumour development. However, as genomic and mammographic screening information were not available in the context of this study, the performance of epigenomic predictors in conjunction with polygenic risk scores, family history, and other predictors of breast cancer risk are beyond the scope of this study. We also acknowledge that although we tested the performance of the PAM classifier on a held-out validation set, this does not preclude the need for replication on a larger, independent cohort. Because large-scale longitudinal studies entail significant costs and logistical challenges, similarly designed studies applying RRBS for DNA methylation analyses have been limited. However, progress in similarly designed prospective studies in recent years [26, 27]

could present an opportunity for these challenges to be overcome in the near future.

Secondly, we acknowledge that analysis of high-dimensional omics-derived datasets by machine learning methods could be vulnerable to overfitting. To mitigate this risk, we included feature reduction steps within our analyses and used a nested cross-validation approach to train the classifier models in addition to evaluate their performance on a held-out validation set, against parallel analyses using label-shuffled datasets. As epigenome-wide analyses and machine learning algorithms improve and become more accessible, we could be poised to see the integration of epigenetic signatures in risk stratification and screening protocols, opening new horizons in the fields of diagnostics and risk prediction, and this could prove to play a critical role in overcoming the challenges of bringing a robust epigenetics-based risk prediction tool to the clinic.

## Conclusions

The findings of this study suggest that gradually accumulated DNA methylation patterns in peripheral blood may occur before clinical manifestation of cancer. Further studies of these changes may provide useful markers for risk stratification and, ultimately, personalized cancer prevention.

## Materials and methods

### Study cohort

The present analysis uses a nested case–control study design with samples from the EPIC-Heidelberg study. Detailed information is provided in the Additional file 1. RRBS was conducted on 739 blood samples collected from women who reported breast cancer over the follow-up period ( $n=359$ ) and cancer-free control participants ( $n=380$ ). Matched controls were selected from cancer-free individuals within the cohort and were matched to cases by age at recruitment ( $\pm 5$  years, with the exception of one pair that had an age difference of 9.9 years), menopausal status, and reported use of hormone therapy and/or contraceptives. All study participants provided written informed consent, and ethical approval for the EPIC study was obtained from the institutional review boards of the International Agency for Research on Cancer and local participating centres.

### Reduced representation bisulphite sequencing (RRBS) and data processing

RRBS was performed as previously described [45], based on DNA extracted from buffy coat samples and FFPE tumour as well as adjacent normal samples. RRBS libraries were sequenced using Illumina HiSeq 2000/3000/4000

platforms in a 50-bp single-end configuration. RRBS data were processed as previously described [45], using a custom pipeline based on Pypiper (v0.6) (<http://code.databio.org/pypiper/>) and Looper (v0.6) (<http://code.databio.org/looper/>). Exploratory analyses were conducted using workflows implemented in RnBeads [46]. Data presented consist of samples that have passed all quality control steps.

### Differential DNA methylation analysis

Differential DNA methylation analyses were conducted for buffy coat samples separately using the output from RnBeads with a custom bioinformatics pipeline [47]. Differences in DNA methylation profiles between cases and controls were identified using a linear model as implemented in the R/Bioconductor package limma [48, 49], with paired analyses, to account for the paired structure of the matched case–control study [46]. Batch correction was conducted on  $M$ -values using surrogate variable analysis as previously described [50]. Models were further adjusted for sequencing lane and length of time to diagnosis. The Enrichr gene list enrichment analysis tool was used to query the GO Biological Process 2021, GO Cellular Component 2021, GO Molecular Function 2021, Reactome 2022, and KEGG 2021 databases for pathway enrichment analysis of the identified DMRs [51, 52]. The annotatr package [53] was used to map DMRs and all analysed regions to genomic contexts as defined in the TxDb.Hsapiens.UCSC.hg38.knownGene and org.Hs.eg.db packages. DMRs were converted to hg19 regions using the liftOver function in the rtracklayer package and mapped to enhancer regions identified in GM12878 through the FANTOM5 project in annotatr [54].

### Marker selection, classifier training and evaluation

Several machine learning classifiers were implemented on mean-centred data using the R package caret. Mean-centring within matched pairs was carried out to account for the paired structure of the matched case–control study. Each classifier was applied on a subset of DNA markers provided by a backward feature selection method (RFE). The predictive performance of each classifier considered was finally assessed by implementing a fivefold nested cross-validation (CV) over 80% of the samples. The overall best-performing machine learning classifier was tested using a held-out set of 68 matched pairs, which were not used in the cross-validation and model development stages.

Full descriptions of the methods are provided in the Additional file 1.

## Abbreviations

AUC	Area under the receiver operating characteristic
CpG	Cytosine-guanine oligodeoxynucleotide
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA elements
EPIC	European Prospective Investigation into Cancer and Nutrition
FDR	False discovery rate
PAM	Prediction analysis for microarrays
RFE	Recursive feature elimination
ROC	Receiver operating characteristic
RRBS	Reduced representation bisulphite sequencing
t-SNE	t-Distributed stochastic neighbour embedding

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13148-023-01509-6>.

**Additional file 1.** Supplementary Methods, Supplementary Tables and References.

**Additional file 2: Fig. S1.** Distribution of participants within the model development and held-out sample sets by age at recruitment, body mass index, exit age, and proportion-of-whole graphs illustrating the distribution of participants by tumour subtype, menopausal status at recruitment, hormonal contraceptive use, hormone therapy use, and pregnancy history.

**Additional file 3. Supplementary Table 3.** List of significantly differentially methylated regions (FDR < 0.05, group mean difference > 0.075) between cases and matched controls in prospectively-collected buffy-coat samples. The correlation between DNA methylation levels at these regions to time to diagnosis was determined by Kendall rank correlation.

**Additional file 4. Supplementary Table 4.** List of significantly differentially methylated regions (FDR < 0.05, group mean difference > 0.075) between cases and matched controls in prospectively-collected buffy-coat samples, limited to samples collected from women diagnosed after the age of 50.

**Additional file 5. Supplementary Table 5.** Top GO Biological Processes, GO Cellular Components, GO Molecular Functions, Reactome, and KEGG Pathways enriched from significantly differentially methylated regions (FDR < 0.05, group mean difference > 0.075) identified when comparing between cases and matched controls in prospectively-collected buffy-coat samples.

**Additional file 6. Supplementary Table 6.** Top GO Biological Processes, GO Cellular Components, GO Molecular Functions, Reactome, and KEGG Pathways enriched from significantly differentially methylated regions (FDR < 0.05, group mean difference > 0.075) identified when overlapping DMRs identified in the main analysis with DMRs identified when samples were limited to participants aged 50 and above at recruitment.

**Additional file 7: Fig. S2.** Relative proportions of hypermethylated, hypomethylated and all regions of the dataset when annotated by Enhancer status as annotated in the FANTOM5 enhancer atlas for the GM12878 human lymphoblastoid cell line; and dngenic annotations.

**Additional file 8. Supplementary Table 7.** Genomic regions in which DNA methylation is significantly correlated to length of time to diagnosis (days) by Kendall rank correlation (FDR < 0.05).

**Additional file 9. Supplementary Table 9.** Genomic regions utilized by the PAM prediction model.

**Additional file 10: Fig. S3.** ROC curves for the tested classifiers. Individual ROC curves are shown for each cross-validation fold. SVM: support vector machines; PLR: penalized logistic regression; NNET: neural network; RF: random forests; LogitBoost: boosted logistic regression; KNN: k-nearest neighbours; PAM: Prediction Analysis for Microarrays; RPART: classification and regression tree

## Acknowledgements

We thank Elizabeth Page and Karen Muller for editing the manuscript.

## Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

## Author contributions

ZH, RK, and CB contributed to conceptualization; CB, RK, ZH, and AR contributed to methodology; SD and RK contributed to investigation—sample collection; AN, SM, CC, AS, BE, PD, ML, JJ, TJ, and VC contributed to investigation—coordination of sample collection, sample preparation, library preparations, sequencing and data preparation; FFC, SM, LB, VC, BE, AG contributed to bioinformatics, statistical analyses and interpretation; FFC, ZH, and LB contributed to writing—original draft; all authors contributed to writing—review and editing; VK, SD, FF, CP, CB, RK, and ZH contributed to funding acquisition; JJ, CB, RK, and ZH contributed to resources; CB, RK, ZH, and AR contributed to supervision. All authors read and approved the final manuscript.

## Funding

This work was supported by grants from the Institut National du Cancer (INCa, France) and the European Commission (EC) Seventh Framework Programme (FP7) Translational Cancer Research (TRANSCAN) Framework, the Fondation ARC pour la Recherche sur le Cancer (France) and La Ligue Française contre le Cancer. The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

## Availability of data and materials

Data generated in this manuscript are available upon reasonable request from the corresponding authors to comply with the IARC and DKFZ institute ethics regulations to protect patient privacy. All requests will be promptly reviewed to verify if request is subject to any intellectual property or confidentiality obligations. Any data and materials that can be shared will be released subject to a Data Transfer Agreement.

## Declarations

### Competing interests

The authors declare no competing financial interests.

Received: 30 January 2023 Accepted: 20 May 2023

Published online: 12 June 2023

## References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394–424.
- Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. *NPJ Precis Oncol.* 2019;3:3.
- Pashayan N, Pharoah PDP. The challenge of early detection in cancer. *Science.* 2020;368(6491):589.
- Pashayan N, Antoniou AC, Ivanus U, Esserman LJ, Easton DF, French D, et al. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nat Rev Clin Oncol.* 2020;17:687.
- Brait M, Sidransky D. Cancer epigenetics: above and beyond. *Toxicol Mech Methods.* 2011;21(4):275–88.
- Zheng Y, Joyce BT, Colicino E, Liu L, Zhang W, Dai Q, et al. Blood epigenetic age may predict cancer incidence and mortality. *EBioMedicine.* 2016;5:68–73.
- Perna L, Zhang Y, Mons U, Hollecsek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenet.* 2016;8:64.

8. Ambatipudi S, Horvath S, Perrier F, Cuenin C, Hernandez-Vargas H, Le Calvez-Kelm F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer*. 2017;75:299–307.
9. Durso DF, Bacalini MG, Sala C, Pirazzini C, Marasco E, Bonafé M, et al. Acceleration of leukocytes' epigenetic age as an early tumor and sex-specific marker of breast and colorectal cancer. *Oncotarget*. 2017;8(14):23237–45.
10. Levine ME, Hosgood HD, Chen B, Absher D, Assimes T, Horvath S. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging*. 2015;7(9):690–700.
11. Yang Z, Wong A, Kuh D, Paul DS, Rakyanc VK, Leslie RD, et al. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol*. 2016;17(1):205.
12. Fransquet PD, Wrigglesworth J, Woods RL, Ernst ME, Ryan J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clin Epigenet*. 2019;11(1):62.
13. Flanagan JM, Munoz-Alegre M, Henderson S, Tang T, Sun P, Johnson N, et al. Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients. *Hum Mol Genet*. 2009;18(7):1332–42.
14. Iwamoto T, Yamamoto N, Taguchi T, Tamaki Y, Noguchi S. BRCA1 promoter methylation in peripheral blood cells is associated with increased risk of breast cancer with BRCA1 promoter methylation. *Breast Cancer Res Treat*. 2011;129(1):69–77.
15. Al-Moghrabi N, Nofel A, Al-Yousef N, Madkhali S, Bin Amer SM, Alaiya A, et al. The molecular significance of methylated BRCA1 promoter in white blood cells of cancer-free females. *BMC Cancer*. 2014;14(1):830.
16. Yang R, Stöcker S, Schott S, Heil J, Marme F, Cuk K, et al. The association between breast cancer and S100P methylation in peripheral blood by multicenter case-control studies. *Carcinogenesis*. 2017;38(3):312–20.
17. Yang R, Pfützke K, Zucknick M, Sutter C, Wappenschmidt B, Marme F, et al. DNA methylation array analyses identified breast cancer-associated HYAL2 methylation in peripheral blood. *Int J Cancer*. 2015;136(8):1845–55.
18. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genom Biol*. 2017;18(1):53.
19. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31:745.
20. Shen SY, Singhania R, Fehrer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563(7732):579–83.
21. Nuzzo PV, Berchuck JE, Korthauer K, Spisak S, Nassar AH, Abou Alaiwi S, et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat Med*. 2020;26:1401.
22. Nassiri F, Chakravarthy A, Feng S, Shen SY, Nejad R, Zuccato JA, et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat Med*. 2020;26(7):1044–7.
23. Jurmeister P, Bockmayr M, Seegerer P, Bockmayr T, Treue D, Montavon G, et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci Transl Med*. 2019;11(509):eaaw8513.
24. Xu Z, Sandler DP, Taylor JA. Blood DNA Methylation and Breast Cancer: A Prospective Case-Cohort Analysis in the Sister Study. *JNCI J Natl Cancer Inst*. 2019;112(1):87–94.
25. Widschwendter M, Evans I, Jones A, Ghazali S, Reisel D, Ryan A, et al. Methylation patterns in serum DNA for early identification of disseminated breast cancer. *Gen Med*. 2017;9(1):115.
26. Widschwendter M, Zikan M, Wahl B, Lempiäinen H, Paprotka T, Evans I, et al. The potential of circulating tumor DNA methylation analysis for the early detection and management of ovarian cancer. *Genom Med*. 2017;9(1):116.
27. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat Commun*. 2020;11(1):3475.
28. Bodelon C, Ambatipudi S, Dugué P-A, Johansson A, Sampson JN, Hicks B, et al. Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies. *Breast Cancer Res*. 2019;21(1):62.
29. Chamberlain JA, Dugué P-A, Bassett JK, Milne RL, Joo JE, Wong EM, et al. DNA methylation in peripheral blood and risk of gastric cancer: a prospective nested case-control study. *Cancer Prevent Res*. 2020;14:233.
30. Feng B, Jiang J, Kraus P, Ng J-H, Heng J-CD, Chan Y-S, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol*. 2009;11(2):197–203.
31. Buganim Y, Faddah Dina A, Cheng Albert W, Itskovich E, Markoulaki S, Ganz K, et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*. 2012;150(6):1209–22.
32. Gallagher KM, Roderick JE, Tan SH, Tan TK, Murphy L, Yu J, et al. ESRRB regulates glucocorticoid gene expression in mice and patients with acute lymphoblastic leukemia. *Blood Adv*. 2020;4(13):3154–68.
33. Wang X, Zhang T, Zhang S, Shan J. Prognostic values of F-box members in breast cancer: an online database analysis and literature review. *Biosci Rep*. 2019;39:1.
34. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G. Pan-Cancer landscape of aberrant DNA methylation across human tumors. *Cell Rep*. 2018;25(4):1066–80.e8.
35. Brinkman AB, Nik-Zainal S, Simmer F, Rodríguez-González FG, Smid M, Alexandrov LB, et al. Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat Commun*. 2019;10(1):1749.
36. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*. 2013;23(3):555–67.
37. de Almeida BP, Apolônio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer*. 2019;19(1):219.
38. Dumeaux V, Ursini-Siegel J, Flatberg A, Fjosne HE, Frantzen J-O, Holmen MM, et al. Peripheral blood cells inform on the presence of breast cancer: a population-based case-control study. *Int J Cancer*. 2015;136(3):656–67.
39. Gao H, Gao R, Zhang L, Xiu W, Zang R, Wang H, et al. Esrrb plays important roles in maintaining self-renewal of trophoblast stem cells (TSCs) and reprogramming somatic cells to induced TSCs. *J Mol Cell Biol*. 2018;11(6):463–73.
40. Latos PA, Goncalves A, Oxley D, Mohammed H, Turro E, Hemberger M. Fgf and Esrrb integrate epigenetic and transcriptional networks that regulate self-renewal of trophoblast stem cells. *Nat Commun*. 2015;6(1):7776.
41. Meng X, Liu X, Guo X, Jiang S, Chen T, Hu Z, et al. FBXO38 mediates PD-1 ubiquitination and regulates anti-tumour immunity of T cells. *Nature*. 2018;564(7734):130–5.
42. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. *Clin Epigenetics*. 2019;11(1):125.
43. Bauer M. Cell-type-specific disturbance of DNA methylation pattern: a chance to get more benefit from and to minimize cohorts for epigenome-wide association studies. *Int J Epidemiol*. 2018;47(3):917–27.
44. Hicks SC, Irizarry RA. methylCC: technology-independent estimation of cell type composition using differentially methylated regions. *Genome Biol*. 2019;20(1):261.
45. Klughammer J, Kiesel B, Roetzer T, Fortelny N, Nemc A, Nenning K-H, et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat Med*. 2018;24(10):1611–24.
46. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 20: comprehensive analysis of DNA methylation data. *Genome Biol*. 2019;20(1):55.
47. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316–9.
48. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat*. 2016;10(2):946–63.
49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl Acids Res*. 2015;43(7): e47.
50. Perrier F, Novoloaca A, Ambatipudi S, Baglietto L, Ghantous A, Perduca V, et al. Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenet*. 2018;10(1):38.
51. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf*. 2013;14:128.

52. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44(W1):W90–7.
53. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. *Bioinformatics.* 2017;33(15):2381–3.
54. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014;507(7493):455–61.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

