Pre-diagnosis blood DNA methylation profiling of twin pairs discordant for breast cancer points to the importance of environmental risk

Authors

Hannes Frederik Bode[1], Liang He[2], Jacob Hjelmborg[2], Jaakko Kaprio[1], Miina Ollikainen[1,3]

1) Institute for Molecular Medicine Finland, University of Helsinki, Tukholmankatu 8, 00290 Helsinki, Finland
2) Department of Public Health, University of Southern Denmark, Winslowvej 9, 5000 Odense, Denmark
3) Minerva Foundation Institute for Medical Research, Tukholmankatu 8, 00290 Helsinki, Finland

Corresponding Author:

Hannes Frederik Bode, hannes.bode@helsinki.fi

Abstract

Background: Breast cancer risk assessment typically relies on consideration of mammography, family history, reproductive history, in addition to assessment of major mutations. However, quantifying the impact of environmental factors, such as lifestyle, can be challenging. DNA methylation (DNAm) presents a promising opportunity for additional information, as it captures effects from both genetics and environment. Previous studies have identified associations and predicted the risk of breast cancer using DNAm in blood, however, these studies did not distinguish between genetic and environmental influences.

Results: Pre-diagnosis blood samples were obtained from 32 monozygotic (MZ)and 76 dizygotic (DZ) female twin pairs discordant for breast cancer. DNAm was profiled using Illumina 450k and EPIC platform. Samples were collected at the mean age of 56 years (standard deviation (SD) =9.90), while the mean age at diagnosis was 66.76 years (SD=6.64). The mean age for censoring controls was 75.20 years (SD=9.45). To identify individual DNAm sites (Cytosine-phosphate-Guanine, CpG) and differentially methylated regions (DMRs) associated with breast cancer risk, survival analysis using paired Cox proportional hazard modeling was performed. Due to the paired modeling, shared genetic and environmental effects, age at entry, and age at sampling were controlled for.

We identified 212 CpGs ($p<6.4*10^{-8}$) and 15 DMRs associated with breast cancer risk across all pairs, with three DMRs overlapping with the individual CpGs. All but one of the 212 CpGs had lower DNAm in cases, suggesting a prevailing trend of hypomethylation of blood DNA prior to breast cancer diagnosis. Altogether 197/212 significant CpGs were also differentially methylated within the 32 MZ twin pairs discordant for breast cancer, suggesting that DNAm at these CpGs is likely independent of genetic effects. Prior research suggests that estrogen regulates at least five of the top CpGs identified. Hence, methylation at these sites may reflect individual differences in estrogen exposure.

Conclusion: In conclusion, most of the identified CpGs associated with future breast cancer diagnosis appear to be independent of genetic effects. This suggests that DNAm could potentially serve as a valuable biomarker for environmental risk factors for breast cancer and may offer potential benefits as a complementary tool to current risk assessment procedures.

Keywords

DNA methylation, breast cancer, discordant twin pairs, survival modelling

Background

Breast cancer (BC) risk assessment tools rely on physiological or genetic screening methods such as mammography, family history and information on reproductive history. In addition to assessment of major mutations, polygenic risk scores are increasingly used to assess overall genetic risk. However, accurately quantifying the impact of environmental factors, including health-related behaviors and occupational exposures, can be challenging, and these factors are often not included in the risk assessment models. In recent years, DNA methylation (DNAm) has emerged as a promising biomarker for BC risk assessment. Therefore, incorporating DNAm information into existing risk assessment models has the potential to improve their accuracy and effectiveness in identifying individuals at increased risk of developing BC.

BC has been linked to DNAm alterations in blood, as evidenced by specific DNAm sites (1–16) and overall average DNAm levels (17) associating with BC. In addition, the studies conducted by Kresovich et al. (2021,(4)), Xiong et al. (2022, (13)) and Chung et al. (2023, (18)) have shown that blood-derived DNAm can be used to predict an individual's risk of developing BC. This presents an opportunity for blood-derived DNAm to serve as a complementary measure to current standard BC risk assessment tools (19,20). However, these predictors, as well as earlier work on DNAm and overall BC risk (2,4,6,11,13,14,16) have not differentiated between genetic and environmental risk factors for BC.

Health-related behaviors and environmental exposures are the main environmental factors that can increase the risk of BC (21), and many of these factors have been shown to affect DNAm as well; e.g. alcohol (22,23), obesity(24,25), physical inactivity(26), hormonal exposure (27), and reproduction-related factors (25,27–29). In addition, genetic variants, including those linked to BC risk, may affect DNAm.

Studies using discordant monozygotic (MZ) twins provide a valuable approach for investigating the impact of environmental factors, e.g., DNAm, on BC risk because the genetic components are fully controlled for in MZ twins, who are genetically identical at the germline. In contrast, genetic effects are partially controlled for in dizygotic (DZ) twin pairs, in which the twins share approximately 50% of their segregating genetic background. Moreover, the twin design effectively controls for all shared environmental influences between twin pairs, regardless of whether they are DZ or MZ twin pairs. Focusing on discordant twin pairs, in which one twin is diagnosed with BC while the other remains cancer-free, helps to minimize the influence of confounding factors. Therefore, the discordant twin design, functioning as a case-control design, dramatically boosts the statistical power and provide more accurate estimates to explore the association between BC risk and DNAm by controlling for shared environmental factors and genetic background. We can then attribute the observed associations to non-shared environmental factors, or in other words, within pair differences in exposure to known or unknown environmental risk factors for BC.

In this study, our objective is to assess the potential impact of DNAm as a biomarker on environmental BC risk. To accomplish this, we leverage a BC discordant twin cohort, utilizing DNAm data collected prior to BC diagnosis in the Finnish Twin Cohort sample. Additionally, we aim to validate our findings by examining an independent BC discordant twin dataset from the Danish Twin Study.

Results

Discovery

In the discovery Model 1 (MZ and DZ pairs), 212 DNAm (Cytosine-phosphate-Guanine, CpG) sites were significantly associated ($p < 6.4*10^{-8}$) with BC (Supplementary table 1, Figure 1A). Among these CpG sites, all except one (cg00550725, in gene *CPNE3*) showed negative association, as indicated by Hazard Ratio (HR) below one. The BC associated hypomethylated CpG sites had HRs ranging from 0.01 to 0.49, while the hypermethylated CpG site had an HR of 3.07. *TDRD1* was the only gene with two significant BC associated CpG sites (cg14779973 and cg27547703).

In addition to the 212 individual CpG sites, 15 DMRs significantly associated with BC in Model 1 (Supplementary table 2). Among these, three DMRs (in genes *SCMH1*, *PXDNL* and *GNAS/RP1-309F20.3*) contain CpG sites that are also significant as single hits in Model 1. Out of the 15 DMRs, 14 exhibit an average HR<1 and lower DNAm in the cases.

Validation in MZ data and comparison between the MZ and DZ model

To explore the significance of the findings in relation to environmental factors, we sought to validate the significant CpG sites from Model 1 using only the MZ twin pairs (Model 2). MZ pairs are fully matched for genetic background, but the sample size is smaller. Model 2 was fitted on the 212 significant CpG sites, hereby 197 CpG sites (93%) meet the Benjamini-Hochberg FDR for the association with BC (Supplementary table 1, Figure 1B). This indicates that the majority of the identified CpG sites are genuinely associated with environmental (non-genetic) risk factors for BC.

Model 3 (containing DZ twin pairs only) was also fitted for the 212 significant CpG sites identified, all passed the Benjamini-Hochberg FDR. Additionally it is to note that the 212 CpG sites had higher effect sizes in the Model 2 compared to Model 3 (regression coefficient 1.31, $p = 0.001$), suggesting that a higher level of genetic matching is associated with a stronger observed effect size (Supplementary table 1, Figure 1C).

Sensitivity analyses

Two sensitivity analyses were done (Supplementary Table 3). The 212 significant CpG sites have Harrell's C indices ranging from 0.59 to 0.72 (mean= 0.65, standard deviation (SD) = 0.02), i.e. notably higher than the expected value of 0.5 for null effects. The E-values were high (E-value of 5.59 for cg00550725) or close to theoretical limits (for HRs <1), indicating that unknown or unmeasured covariates are unlikely to account for the association between the CpG sites and BC.

Comparison of the significant CpG sites with the Danish Twin study

The Finnish and Danish datasets shared 98 out of the 212 CpG sites identified in the Finnish analysis. The remaining sites were unavailable due to platform differences, with the Danish data solely based on the 450K platform. Out of these shared CpG sites 22 had the same effect direction in the Finnish data (Model 1 and Model 2) and the Danish data (Model 2R). However, the Danish data (Model 2R) did not yield any replication of Finnish data (Model 1 or Model 2) under Benjamini-Hochberg FDR (Supplementary table 4, Figure 1D).

Discussion

In this study 212 CpG sites and 15 DMRs associated with BC using a discordant twin design, matching for familial confounders. Among these DMRs are three that contain CpG sites with an individually significant p-value. These DMRs are in the genes *SCMH1*, *PXDNL*, and *GNAS/RP1-309F20.3*. Two individually significant CpG sites were found to be located in the gene *TDRD1*. The majority of the CpG sites (197 out of 212) were identified also in the model using MZ pairs only, which is fully matched for genetic confounding. This implies that these 197 CpG sites associate with BC independent of genetic factors and are likely attributed to environmental effects. In addition, significantly higher effect sizes were observed across the 212 CpG sites among the MZ pairs, compared to DZ pairs. No CpG sites was replicated in the independent dataset from the Danish twin study.

DNAm may change at specific CpG sites due to exposure to a BC risk factor. Several environmental and health related behavioral factors, such as alcohol consumption (22,23), obesity (24,25), physical inactivity (26), hormonal exposure (27), hormonal disruption (30), and multiple reproduction-related factors (25,27–29) have been found to associate with DNAm. It is important to note that these are also known risk factors for BC (21). Based on this, the association between DNAm and specifically environmental BC risk which was observed in this study could be hypothesized as a summary of exposure to these factors. Hereby, the twin with BC has potentially been exposed to a risk factor with greater extent compared to her co-twin who has

not had BC, leading to within-pair difference in DNAm. However, determining the specific contribution of individual factor is not possible within the scope of this study. Interestingly, we identified several genes (*TDRD1*, *SCMH1*, *PXDNL* and *GNAS*), each containing multiple sites associating positively with BC, which have been shown to be regulated by estrogen (see Textbox 1). In addition, the only gene (*CPNE3*) that has a significantly hypermethylated CpG site in cases in our study, has been indicated to be under the regulation of estrogen in BC (Textbox 1). Identifying potentially estrogen regulated genes exhibiting differential methylation in relation to environmental factors supports our hypothesis that DNAm can summarize environmental exposure or disruption of hormones, which are known risk factors for BC (21).

Textbox 1: Selected Genes with high relevance for environmental risk for breast cancer and indication for estrogen dependent regulation.

Gene	Description
TDRD1	*TDRD1* is regulated by estrogen-regulated gene (*ERG*) in prostate (31,32) and colon cancer (33). *ERG*, which overexpressed in the presence of estrogen (34), promotes demethylation of *TDRD1* at the promoter site, leading to increased expression of *TDRD1* (32).
SCMH1	Methylation of the *SCMH1* gene is reduced with increased lifetime estrogen exposure (27) and is overexpressed in women with polycystic ovary syndrome (35), a condition associated with dysregulated estrogen biology (36).
PXDNL	*PXDNL* is associated with poor survival in BC and therefore is suggested as a potential prognostic biomarker for BC (37). In prostate cancer cell line expression of PXDN, a homolog of *PXDNL*, is downregulated by *Estrogen Related Receptor Beta (ERR Beta)* (38). *ERR* family members are known to downregulate genes that are otherwise regulated by estrogen (39).
GNAS	*GNAS* is down regulated by estrogen in an ovariectomized rat model (40).
CPNE3	*CPNE3* has been found to be upregulated by estrogen in BC (41). Additionally, studies have shown that *CPNE3* expression is higher in BC cell lines that are resistant to endocrine treatment (42), compared to cell lines that are responsive to estrogen (43,44). Endocrine treatment resistance in BC is characterized by a state where external estrogen signaling is not necessary for the tumor to grow, but receptor activity itself can remain essential. This resistance can be achieved through cell internal changes, such as mutations or epigenetic remodeling, which enable the activation of the estrogen receptor independent of external estrogen stimulation (45), leading potentially higher *CPNE3* expression.

Altogether 68 CpG sites identified in the current study Model 1 are located in the same genes that have been previously associated with BC through DNAm (2,4,6,11,13,14,16) (Supplementary table 5). In six of these seven studies, blood samples were collected before breast cancer diagnosis (average time to diagnosis reported only for five studies: 1.3 years (14), 3.8 years (11), 5.2 years (4), 5.6 years (6), and 7.2 years (2)), while one study obtained samples at the time of diagnosis, but prior to treatment (16). Two studies focused on case cohorts using time to diagnosis as a variable (4,11), while the other studies included both cases and controls (2,6,13,14,16). In each of the seven studies, the focus was on overall breast cancer risk, not distinguishing between genetic and environmental risk factors.

Among these 68 CpG derived from these seven studies, four CpG sites passed Benjamini-Hochberg FDR in the Finnish data of Model 1 (Supplementary table 6), with cg21769444 being one of the 212 discovery CpG sites. The effects sizes reported in the literature vary due to differences in analysis approaches, including fold change in DNAm between cases and controls, Hazard Ratio, and differences in DNAm between cases and controls. Notably, cg01259104/*ANKLE2*, cg04248461/*DIP2C*, and cg21769444/*NUDT3* showed similar effect directions as in this study, while cg05375728/*DAB1* exhibited an effect in the opposite direction.

Further, Chung et al. (2023, (18)) conducted a prospective EWAS on blood samples for breast cancer (BC), identifying 187 DMRs associated with future BC diagnosis. None of these DMRs replicated in our study, however, four individual CpG sites within five different DMRs replicated under Benjamini-Hochberg FDR in our Model 1 (Supplementary table 7).

Strengths and limitations

A notable strength of this study is that the DNAm data used for analysis was collected on average, almost 11 years prior to BC diagnosis, showing that DNAm associates with BC already well before the cancer is diagnosed. Hereby, DNAm has potential for identifying individuals at a high risk of developing BC at an early stage, which could be used to implement proactive interventions, such as more frequent screenings or targeted preventive measures. A further strength of our study lies in its comprehensive approach, accounting for both known and unknown factors that could confound DNAm analysis, including age, technical variation, and familial effects. Additionally, the sampling strategy implemented prior to diagnosis is a crucial advantage, as it effectively minimizes the potential influence of breast cancer as a disease and its treatment, thus significantly enhancing the reliability and validity of our findings.

Nevertheless, several potential limitations in this study should be noted. Firstly, the absence of data on BC subtypes could be crucial in refining the outcomes. Distinct BC subtypes require different treatments and have varying susceptibilities to environmental risk factors, mainly in the case of as hormone receptor-positive versus hormone receptor-negative BC (42). Secondly, the sample size of MZ twin pairs is small,

which reduces statistical power. The inclusion of DZ twins can partly address this issue; however, it introduces bias based on genetic factors. Thirdly, replication in the Danish data may not be optimal due to differences in sample size and age at sampling. However, these limitations are beyond control as they depend on the data availability.

Conclusion

We demonstrate the presence of DNAm alterations in blood on average over 11 years before the actual BC diagnosis, likely independent of familial factors, like shared early life environment and germline genetics. Individual environmental exposures or de novo mutations are likely contributing factors that could explain this observation. The study identified associations between BC risk and DNAm of genes involved in BC biology, specifically of estrogen-related genes. Furthermore, our findings suggest that DNAm could be a promising addition to BC risk assessment toolset for identifying individuals who have a higher likelihood of developing BC due to environmental experiences and exposures. Our findings warrant future investigations in much larger prospective cohorts to clarify which environmental factors are most relevant, and associate with BC risk through DNAm.

Methods

The Finnish Twin Cohort

The Finnish Twin Cohort, consisting of individuals born before 1958, was established in 1975, recruitment was completed by May 1, 1976, and the followed-up period lasted until December 31, 2018. To obtain cancer diagnosis data during the study period, the Finnish Twin Cohort was linked to the Finnish Cancer registry. Information on death and emigration was obtained from the Finnish population registry. In the 1990s, blood samples were collected from a subset of individuals and DNA was extracted and stored in the Biobank of the Finnish Institute for Health and Welfare. DNAm data was subsequently generated from these samples.

A group of 108 pairs of female twins who showed discordance for BC at the end of the follow-up period were selected from among all pairs with DNA from the Finnish Twin Cohort 32 pairs were MZ, and 76 pairs were DZ (Table 1). Among the cases, BC was either their first or only cancer diagnosis, while the controls remained cancer-free during the follow-up. The follow-up period was considered to end for cases at the time of diagnosis and for controls either at death (n=18) or latest at the end of the study in 2018 (n=90).

The average age of study entry was 33.78 years (SD = 9.53 years) and the average age for blood sample collection was 56.00 years (SD = 9.90 years). For cases, the age at diagnosis was on average 66.76 years (SD = 6.64 years), and for controls, the end of follow-up was at an average age of 75.20 years (SD = 9.45 years). The mean time between blood sampling and diagnosis was 10.76 years (SD = 6.65).

Data on epidemiological risk factors for BC were obtained from a health-related questionnaire collected in 1975 (Table 2), while information on the number of children and age at first birth was obtained from the Finnish Population Register (Table 2) (46). The association between these variables and breast cancer discordance was examined using conditional logistic regression. None of these variables showed a significant association with BC discordance (Supplementary Table 8).

The Danish Twin Study

The Danish sample is selected from two Danish twin cohorts, including the Longitudinal Study of Aging in Danish Twins (LSADT) study and the Middle Age Danish Twins (MADT) study. Details about these twin cohorts are described previously in (47,48). We only included those twins that have both DNAm (1180 samples) and the information about BC diagnosis, which was retrieved through the link between the twin registry and the cancer registry in the NorTwinCan database (49). Eighty-six twins in LSADT were measured twice in 1997 and 2007, respectively, for their DNAm and we included only the early measurement in 1997 to increase the sample size for the analysis of pre-diagnosis DNAm. Among these twins, 11 twin pairs (eight MZ and three DZ pairs) that are discordant for BC and of whom the methylation was measured prior to the diagnosis were included in the analysis. The mean age at the diagnosis of these 11 twin pairs is 78.09 (SD=9.62 years) and the mean age at the DNAm profiling is 71.32 (SD=6.61 years). The end of follow-up was defined the same way as for the Finnish dataset.

DNA methylation data

Among the 108 pairs from the Finnish Twin Cohort, DNAm was measured for four pairs using the Illumina Infinium HumanMethylation450 (450k) platform and for 104 pairs the Illumina Infinium Methylation (EPIC) platform (Illumina, San Diego, CA, USA), (Table 1). Twins in a pair shared the same platform technology and were sampled and processed at the same time. Preprocessing of the DNAm data was done using the *meffil* R package (50). Sample quality was assessed based on the following criteria: (1) mean difference between X and Y (technical noise in female samples) chromosome signals was less than -2, (2) mean methylated signal did not deviate from the regression line (mean methylated signal linear regressed over mean unmethylated signal) by more than three standard deviations, (3) sample was not an outlier based on Illumina control probes, (4) percentage of probes with only background signal was less than 20, and (5) percentage of

probes with less than three beads was less than 20. All samples that met all the above criteria passed the quality control. In addition, ambiguous mapping and poor-quality probes based on Zhou et al. (2017, (51)) and probes binding to sex chromosomes were removed to address ambiguity of the DNAm signal based on X-chromosome inactivation (52,53).
Following the quality control, the DNAm data underwent preprocessing in two separate batches due to use of two different types of microarray platforms, EPIC and 450K. The CpG sites shared across the platforms were preprocessed together, while the EPIC specific CpG sites were preprocessed separately using the same approach.

The preprocessing was performed by functional normalization using the first 15 principal components of the control probes, to eliminate unwanted technical variation using the *meffil* R package (50). Additionally, to reduce probe bias, beta-mixture quantile normalization (54) using the *wateRmelon* R package (55) was applied. Next, the CpG probes exclusively present on the EPIC platform were merged with the set of probes common between the EPIC and 450k platforms. Finally, the beta values were scaled based on the standard deviation of each CpG probe across all samples.

After performing quality control and preprocessing, a total of 52333 probes were removed due to their insufficient quality, and 9918 probes were removed due to their binding to sex chromosomes. A final set of 778861 probes were retained, consisting of 336849 probes (43%) shared by the 450k and EPIC data and 442012 probes exclusive to the EPIC platform.

In the Danish twin cohorts, DNAm was measured from the buffy coat samples stored at -80°C in 24 hours after the blood was collected using the HumanMethylation450 (450k) platform (Illumina, San Diego, CA, USA). Quality control for sample and probe exclusion were conducted with the *MethylAid* (56) and *Minfi* (57) R packages. More detailed steps of quality control and criteria for excluding samples and probes are described in the previous study (58). After further excluding any CpG sites that had a missing rate of >10% across the whole 1180 samples, 451,471 CpG sites remained in the survival analysis.

Survival modelling

To investigate the association between DNAm and BC risk, a survival analysis was conducted using Cox proportional hazard regression models in the R package *survival* (59). To ensure that all CpG sites meet the model assumptions, the proportional hazard assumption of time independence of the hazard for the within-pair difference in methylation beta value was tested using the *cox.zph()* function, and Schoenfeld residuals were examined for the significant CpG sites. For the analysis, four different models with the same survival term and covariates , but different test population were run (Equation 1).

Equation 1: Survival term (age at censoring | BC status) ~ within-pair difference in methylation beta value + pairwise mean in methylation beta value + frailty term of pair identifier.

The survival term in this model refers to the survival outcome observed during the follow-up period. To adjust for the methylation levels specific to each twin pair, the mean beta values for each pair were included as a covariate. The pair identifier is considered as a random effect. Additionally, based on the study design the pairs were matched on potential confounding variables, such as chip platform, age at entry, age at sampling, sex ,and early life environment.

The primary analysis involved 108 twin pairs that were sampled before BC diagnosis (Model 1). This was considered as the discovery model, with the significance threshold defined as Bonferroni $p < 6.4*10-8$, which corresponds to a p-value of 0.05 divided by the number of CpG sites (n = 778861). The statistically significant CpG sites were validated in zygosity-specific analysis of MZ pairs (n=32, Model 2) and DZ pairs (n=76, Model 3), to assess the role of genetic vs environmental effects among the significant methylation sites. The resulting p-values from the follow-up models (Model 2 and 3) were corrected by Benjamini-Hochberg procedure, and the significance was set to FDR < 0.05.

To replicate our significant findings in an independent twin sample, a comparable survival model was fitted using the Danish data of 11 twin pairs discordant for BC (Model 2R). The resulting p-values were corrected by Benjamini-Hochberg procedure, and the significance was set to FDR < 0.05.

Sensitivity analyses

To assess the robustness of the identified significant DNAm sites, a sensitivity analysis was conducted by performing a 10-fold cross-validated calculation of the Harrell's Concordance index for within-pair difference in DNAm at the significant CpG sites. These results were compared with the concordance of the fitted covariates (60).

To evaluate the possible effect of unmeasured confounding variables on the association between DNAm and survival outcomes, E-values were computed for each significant CpG site using the *Evalue* R package (61,62). To simplify the E-values comparison of CpG sites with different effect (HR) directions, E-values are represented in a simplified manner. For CpG sites with HR<1, the E-values are provided as a fraction with the numerator set as 1.

DMR analysis

Differentially methylated regions (DMR) were identified using the *ipDMR()* function (63) of the *ENmix* R-package (64). Hereby neighboring CpG sites within a maximum distance of 500 base pairs were identified, and their combined p-values were calculated using the p-values derived from Model 1. Using a Benjamini-Hochberg FDR < 0.001, all significant pairs of CpG sites were selected. These significant CpG pairs were then merged into broader regions using an approach like single linkage clustering, where neighboring CpG sites within a maximum distance of 500 base pairs were linked together. The association between each broader region and the survival outcome was then calculated by summarizing the associations of all CpG sites present in such region, based on the individual CpG sites p-values derived from Model 1. A DMR was considered significant if it contained three or more CpG sites with unidirectional methylation association and had a Benjamini-Hochberg FDR < 0.001 combined across all CpG sites.

List of abbreviations

BC: breast cancer

CI: confidence interval

CpG: Cytosine-phosphate-Guanine

DNAm: DNA methylation

DZ: dizygotic

FDR: false discovery rate

HR: Hazard Ratio

LSADT: Longitudinal Study of Aging in Danish Twins Study

MADT: Middle Age Danish Twins Study

MZ: monozygotic

OR: Odds Ratio

SD: standard deviation

Declarations

Ethics approval and consent to participate

Informed consent was obtained before the beginning of the studies in 1975, and upon every contact with the study subjects. When clinical investigations were undertaken with sampling of biological material, written informed consent was obtained. Ethics approval of these procedures was provided in multiple studies, the last one on the transfer of biological samples to the THL Biobank by the Hospital District of Helsinki and Uusimaa ethics board in 2018.

The two Danish twin studies (LSADT and MADT) were approved by the Regional Committees on Health Research Ethics for Southern Denmark (S-VF-19980072), and written informed consents were obtained from all participants.

Availability of data and materials

The Finnish Twin Cohort datasets utilized in the current study are stored in the Biobank of the Finnish Institute for Health and Welfare, Helsinki, Finland. The data is publicly available for use by qualified researchers through a standardized application procedure (https://thl.fi/en/web/thl-biobank/forresearchers).

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the European Union's Horizon 2020 Research and Innovation Programme, Marie Skłodowska-Curie (JK, grant number 859860). This project has received further funding from the Academy of Finland (MO, grant numbers 297908, 328685 & JK, grant 336823) and the Sigrid Juselius Foundation (MO and JK).

Authors' contributions

All authors contributed to conceptualization and methodology. HB and LH contributed to the bioinformatics and statistical analyses. HB, MO and JK contributed to the interpretation. HB contributed to writing of the original draft; all authors contributed to reviewing and editing of the manuscript. JH, LH, JK and MO contributed to supervision of the project. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the participants for their invaluable contribution to the study. The technical staff at the Finnish Twin Cohort stud and the Danish twin studies are acknowledged for their help in collecting the data. The authors also wish to thank Mia Urjansson and Teemu Palviainen, from the University of Helsinki, for their valuable help with the data collection.

References

1. Johansson A, Flanagan JM. Epigenome-wide association studies for breast cancer risk and risk factors. Trends Cancer Res [Internet]. 2017 [cited 2022 Apr 7];12:19. Available from: /pmc/articles/PMC5612397/

2. Massi MC, Dominoni L, Ieva F, Fiorito G. A Deep Survival EWAS approach estimating risk profile based on pre-diagnostic DNA methylation: An application to breast cancer time to diagnosis. PLoS Comput Biol [Internet]. 2022 Sep 1 [cited 2023 Aug 2];18(9). Available from: /pmc/articles/PMC9536632/

3. Joo JE, Dowty JG, Milne RL, Wong EM, Dugué PA, English D, et al. Heritable DNA methylation marks associated with susceptibility to breast cancer /631/67/69 /631/337/176/1988 /692/699/67/1347 /692/308/2056 /45 /45/61 article. Nat Commun [Internet]. 2018 Dec 1 [cited 2021 Jun 12];9(1):1–12. Available from: www.nature.com/naturecommunications

4. Kresovich JK, Xu Z, O'Brien KM, Shi M, Weinberg CR, Sandler DP, et al. Blood DNA methylation profiles improve breast cancer prediction. Mol Oncol [Internet]. 2022 Jan 1 [cited 2022 Feb 25];16(1):42. Available from: /pmc/articles/PMC8732352/

5. Tuminello S, Zhang Y, Yang L, Durmus N, Snuderl M, Heguy A, et al. Global DNA Methylation Profiles in Peripheral Blood of WTC-Exposed Community Members with Breast Cancer. Int J Environ Res Public Heal 2022, Vol 19, Page 5104 [Internet]. 2022 Apr 22 [cited 2023 Aug 2];19(9):5104. Available from: https://www.mdpi.com/1660-4601/19/9/5104/htm

6. Ennour-Idrissi K, Dragic D, Issa E, Michaud A, Chang SL, Provencher L, et al. DNA Methylation and Breast Cancer Risk: An Epigenome-Wide Study of Normal Breast Tissue and Blood. Cancers (Basel) [Internet]. 2020 Nov 1 [cited 2023 Aug 2];12(11):1–16. Available from: https://pubmed.ncbi.nlm.nih.gov/33113958/

7. Ho PJ, Dorajoo R, Ivanković I, Ong SS, Khng AJ, Tan BKT, et al. DNA methylation and breast cancer-associated variants. Breast Cancer Res Treat [Internet]. 2021 Aug 1 [cited 2023 Aug 2];188(3):713–27. Available from: https://link.springer.com/article/10.1007/s10549-021-06185-9

8. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet [Internet]. 2013 Mar [cited 2023 Aug 2];22(5):843–51. Available from: https://pubmed.ncbi.nlm.nih.gov/23175441/

9. Scott CM, Wong EM, Joo JHE, Dugué PA, Jung CH, O'Callaghan N, et al. Genome-wide DNA methylation assessment of 'BRCA1-like' early-onset breast cancer: Data from the Australian Breast Cancer Family Registry. Exp Mol Pathol. 2018 Dec 1;105(3):404–10.

10. Anjum S, Fourkala EO, Zikan M, Wong A, Gentry-Maharaj A, Jones A, et al. A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. Genome Med [Internet]. 2014 Jun 27 [cited 2023 Aug 2];6(6). Available from: https://pubmed.ncbi.nlm.nih.gov/25067956/

11. 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 [Internet]. 2020 Jan 1 [cited 2023 Aug 2];112(1):87. Available from: /pmc/articles/PMC7489106/

12. Yang Y, Wu L, Shu XO, Cai Q, Shu X, Li B, et al. Genetically Predicted Levels of DNA Methylation Biomarkers and Breast Cancer Risk: Data From 228 951 Women of European Descent. J Natl Cancer Inst [Internet]. 2020 [cited 2023 Aug 2];112(3):295–304. Available from: https://pubmed.ncbi.nlm.nih.gov/31143935/

13. Xiong Z, Yang L, Ao J, Yi J, Zouxu X, Zhong W, et al. A Prognostic Model for Breast Cancer Based on Cancer Incidence-Related DNA Methylation Pattern. Front Genet [Internet]. 2022 Jan 3 [cited 2023 Aug 2];12. Available from: https://pubmed.ncbi.nlm.nih.gov/35047022/

14. Xu Z, Bolick SCE, Deroo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide Association Study of Breast Cancer Using Prospectively Collected Sister Study Samples. JNCI J Natl Cancer Inst [Internet]. 2013 May 5 [cited 2023 Aug 2];105(10):694. Available from: /pmc/articles/PMC3653821/

15. Parashar S, Cheishvili D, Mahmood N, Arakelian A, Tanvir I, Khan HA, et al. DNA methylation signatures of breast cancer in peripheral T-cells. BMC Cancer [Internet]. 2018 May 18 [cited 2023 Aug 2];18(1). Available from: /pmc/articles/PMC5960123/

16. Cappetta M, Fernandez L, Brignoni L, Artagaveytia N, Bonilla C, López M, et al. Discovery of novel DNA methylation biomarkers for non-invasive sporadic breast cancer detection in the Latino population. Mol Oncol [Internet]. 2021 Feb 1 [cited 2023 Aug 2];15(2):473–86. Available from: https://pubmed.ncbi.nlm.nih.gov/33145876/

17. Tang Q, Cheng J, Cao X, Surowy H, Burwinkel B. Blood-based DNA methylation as biomarker for breast cancer: a systematic review. Clin Epigenetics [Internet]. 2016 Nov 14 [cited 2023 Aug 2];8(1). Available from: /pmc/articles/PMC5109688/

18. Chung FFL, Maldonado SG, Nemc A, Bouaoun L, Cahais V, Cuenin C, et al. Buffy coat signatures of breast cancer risk in a prospective cohort study. Clin Epigenetics [Internet]. 2023 Dec 1 [cited 2023 Aug 2];15(1). Available from: /pmc/articles/PMC10262593/

19. Pashayan N, Pharoah P. Population-based screening in the era of genomics. Per Med [Internet]. 2012 Jun [cited 2023 Aug 2];9(4):451–5. Available from: https://pubmed.ncbi.nlm.nih.gov/22984365/

20. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. J Natl Cancer Inst [Internet]. 2014 Nov 1 [cited 2023 Aug 2];106(11). Available from: https://pubmed.ncbi.nlm.nih.gov/25392194/

21. Kashyap D, Pal D, Sharma R, Garg VK, Goel N, Koundal D, et al. Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. Biomed Res Int [Internet]. 2022 [cited 2023 Aug 2];2022. Available from: /pmc/articles/PMC9038417/

22. Varela-Rey M, Woodhoo A, Martinez-Chantar ML, Mato JM, Lu SC. Alcohol, DNA Methylation, and Cancer. Alcohol Res [Internet]. 2013 [cited 2023 Aug 2];35(1):25. Available from: /pmc/articles/PMC3860423/

23. Mahna D, Puri S, Sharma S. DNA methylation signatures: Biomarkers of drug and alcohol abuse. Mutat Res Mutat Res. 2018 Jul 1;777:19–28.

24. Dragic D, Ennour-Idrissi K, Michaud A, Chang SL, Durocher F, Diorio C. Association Between BMI and DNA Methylation in Blood or Normal Adult Breast Tissue: A Systematic Review. Anticancer Res [Internet]. 2020 Apr 1 [cited 2023 Aug 2];40(4):1797–808. Available from: https://pubmed.ncbi.nlm.nih.gov/32234868/

25. Chen M, Wong EM, Nguyen TL, Dite GS, Stone J, Dugué P-A, et al. DNA methylation-based biological age, genome-wide average DNA methylation, and conventional breast cancer risk factors. Sci Rep [Internet]. 2019 Dec 1 [cited 2021 Jul 28];9(1). Available from: /pmc/articles/PMC6803691/

26. Światowy WJ, Drzewiecka H, Kliber M, Sąsiadek M, Karpiński P, Pławski A, et al. Physical Activity and DNA Methylation in Humans. Int J Mol Sci [Internet]. 2021 Dec 1 [cited 2023 Aug 2];22(23). Available from: https://pubmed.ncbi.nlm.nih.gov/34884790/

27. Johansson A, Palli D, Masala G, Grioni S, Agnoli C, Tumino R, et al. Epigenome-wide association study for lifetime estrogen exposure identifies an epigenetic signature associated with breast cancer risk. Clin Epigenetics [Internet]. 2019 Apr 30 [cited 2023 Aug 2];11(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31039828/

28. Levine ME, Lu AT, Chen BH, Hernandez DG, Singleton AB, Ferrucci L, et al. Menopause accelerates biological aging. Proc Natl Acad Sci U S A [Internet]. 2016 Aug 16 [cited 2022 Feb 25];113(33):9327–32. Available from: /pmc/articles/PMC4995944/

29. Kresovich JK, Xu Z, O'Brien KM, Weinberg CR, Sandler DP, Taylor JA. Methylation-Based Biological Age and Breast Cancer Risk. JNCI J Natl Cancer Inst [Internet]. 2019 Oct 1 [cited 2023 Aug 2];111(10):1051. Available from: /pmc/articles/PMC6792078/

30. Maitre L, Jedynak P, Gallego M, Ciaran L, Audouze K, Casas M, et al. Integrating -omics approaches into population-based studies of endocrine disrupting chemicals: A scoping review. Environ Res. 2023 Jul 1;228:115788.

31. Xiao L, Lanz RB, Frolov A, Castro PD, Zhang Z, Dong B, et al. The Germ Cell Gene TDRD1 as an ERG Target Gene and a Novel Prostate Cancer Biomarker. Prostate [Internet]. 2016 Oct 1 [cited 2023 Aug 2];76(14):1271–84. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pros.23213

32. Kacprzyk LA, Laible M, Andrasiuk T, Brase JC, Börno ST, Fälth M, et al. ERG induces epigenetic activation of Tudor domain-containing protein 1 (TDRD1) in ERG rearrangement-positive prostate cancer. PLoS One [Internet]. 2013 Mar 29 [cited 2023 Aug 2];8(3). Available from: https://pubmed.ncbi.nlm.nih.gov/23555854/

33. Mo HY, Choi EJ, Yoo NJ, Lee SH. Mutational alterations of TDRD 1, 4 and 9 genes in colorectal cancers. Pathol Oncol Res [Internet]. 2020 Jul 1 [cited 2023 Aug 2];26(3):2007–8. Available from: https://pubmed.ncbi.nlm.nih.gov/32036563/

34. Setlur SR, Mertz KD, Hoshida Y, Demichelis F, Lupien M, Perner S, et al. Estrogen-Dependent Signaling in a Molecularly Distinct Subclass of Aggressive Prostate Cancer. JNCI J Natl Cancer Inst [Internet]. 2008 Jun 4 [cited 2023 Aug 2];100(11):815–25. Available from: https://dx.doi.org/10.1093/jnci/djn150

35. Vázquez-Martínez ER, Gómez-Viais YI, García-Gómez E, Reyes-Mayoral C, Reyes-Muñoz E, Camacho-Arroyo I, et al. DNA methylation in the pathogenesis of polycystic ovary syndrome. Reproduction [Internet]. 2019 Jul 1 [cited 2023 Aug 2];158(1):R27–40. Available from: https://rep.bioscientifica.com/view/journals/rep/158/1/REP-18-0449.xml

36. Xu XL, Deng SL, Lian ZX, Yu K. Estrogen Receptors in Polycystic Ovary Syndrome. Cells [Internet]. 2021 Feb 1 [cited 2023 Aug 2];10(2):1–13. Available from: /pmc/articles/PMC7924872/

37. Li Y, Jiao Y, Luo Z, Li Y, Liu Y. High peroxidasin-like expression is a potential and independent prognostic biomarker in breast cancer. Medicine (Baltimore) [Internet]. 2019 Nov 1 [cited 2023 Aug

2];98(44):e17703. Available from: /pmc/articles/PMC6946426/

38. Lu Y, Li J, Cheng J, Lubahn DB. Messenger RNA profile analysis deciphers new Esrrb responsive genes in prostate cancer cells. BMC Mol Biol [Internet]. 2015 Dec 1 [cited 2023 Aug 2];16(1). Available from: https://pubmed.ncbi.nlm.nih.gov/26627478/

39. Tanida T, Matsuda KI, Yamada S, Hashimoto T, Kawata M. Estrogen-related Receptor β Reduces the Subnuclear Mobility of Estrogen Receptor α and Suppresses Estrogen-dependent Cellular Function. J Biol Chem [Internet]. 2015 May 8 [cited 2023 Aug 2];290(19):12332–45. Available from: https://pubmed.ncbi.nlm.nih.gov/25805499/

40. Pham LT, Yamanaka K, Miyamoto Y, Waki H, Gouraud SSS. Estradiol-dependent gene expression profile in the amygdala of young ovariectomized spontaneously hypertensive rats. Physiol Genomics [Internet]. 2022 Mar 1 [cited 2023 Aug 2];54(3):99–114. Available from: https://journals.physiology.org/doi/10.1152/physiolgenomics.00082.2021

41. Harvell DME, Richer JK, Allred DC, Sartorius CA, Horwitz KB. Estradiol regulates different genes in human breast tumor xenografts compared with the identical cells in culture. Endocrinology [Internet]. 2006 Feb [cited 2023 Aug 2];147(2):700–13. Available from: https://pubmed.ncbi.nlm.nih.gov/16239301/

42. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA [Internet]. 2019 Jan 22 [cited 2023 Aug 2];321(3):288–300. Available from: https://jamanetwork.com/journals/jama/fullarticle/2721183

43. Tilghman SL, Townley I, Zhong Q, Carriere PP, Zou J, Llopis SD, et al. Proteomic signatures of acquired letrozole resistance in breast cancer: suppressed estrogen signaling and increased cell motility and invasiveness. Mol Cell Proteomics [Internet]. 2013 Sep [cited 2023 Aug 2];12(9):2440–55. Available from: https://pubmed.ncbi.nlm.nih.gov/23704778/

44. Walker RR, Gallegos KM, Bratton MR, Lemieux KP, Zhang K, Wang G, et al. Acquisition of Letrozole Resistance Through Activation of the p38/MAPK Signaling Cascade. Anticancer Res [Internet]. 2021 Feb 1 [cited 2023 Aug 2];41(2):583–99. Available from: https://pubmed.ncbi.nlm.nih.gov/33517263/

45. Hartkopf AD, Grischke EM, Brucker SY. Endocrine-Resistant Breast Cancer: Mechanisms and Treatment. Breast Care (Basel) [Internet]. 2020 [cited 2023 Aug 2];15(4):347–54. Available from: https://pubmed.ncbi.nlm.nih.gov/32982644/

46. Rose RJ, Latvala A, Silventoinen K, Kaprio J. Alcohol consumption at age 18-25 and number of children at a 33-year follow-up: Individual and within-pair analyses of Finnish twins. Alcohol Clin Exp Res [Internet]. 2022 Aug 1 [cited 2023 Aug 2];46(8):1552–64. Available from: https://pubmed.ncbi.nlm.nih.gov/35719054/

47. Pedersen DA, Larsen LA, Nygaard M, Mengel-From J, McGue M, Dalgård C, et al. The Danish Twin Registry: An Updated Overview. Twin Res Hum Genet [Internet]. 2019 Dec 1 [cited 2023 Aug 2];22(6):499. Available from: /pmc/articles/PMC8039015/

48. Skytthe A, Harris JR, Czene K, Mucci L, Adami H-O, Christensen K, et al. Cancer Incidence and Mortality in 260,000 Nordic Twins With 30,000 Prospective Cancers. Twin Res Hum Genet [Internet]. 2019 Apr 1 [cited 2021 Sep 7];22(2):99–107. Available from: https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/cancer-incidence-and-mortality-in-260000-nordic-twins-with-30000-prospective-cancers/563BA72D2EB383C6384C61F4174796B3

49. Harris JR, Hjelmborg J, Adami HO, Czene K, Mucci L, Kaprio J. The Nordic Twin Study on Cancer — NorTwinCan. Twin Res Hum Genet [Internet]. 2019 Dec 1 [cited 2023 Aug 2];22(6):817–23. Available from: https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/nordic-twin-study-on-cancer-nortwincan/180390D6C3F8A0A0AC623016D84F268A

50. Min JL, Hemani G, Smith GD, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics [Internet]. 2018 Dec 1 [cited 2023 Aug 2];34(23):3983–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29931280/

51. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res [Internet]. 2017 Feb 28 [cited 2021 Jul 28];45(4):e22–e22. Available from: https://academic.oup.com/nar/article/45/4/e22/2290930

52. Inkster AM, Wong MT, Matthews AM, Brown CJ, Robinson WP. Who's afraid of the X? Incorporating the X and Y chromosomes into the analysis of DNA methylation array data. Epigenetics Chromatin [Internet]. 2023 Dec 1 [cited 2023 Aug 2];16(1). Available from: https://pubmed.ncbi.nlm.nih.gov/36609459/

53. Feil R, Khosla S. Genomic imprinting in mammals: An interplay between chromatin and DNA methylation? Trends Genet [Internet]. 1999 Nov 1 [cited 2023 Aug 2];15(11):431–5. Available from: http://www.cell.com/article/S0168952599018223/fulltext

54. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics [Internet]. 2013 Jan 1 [cited 2023 Aug 2];29(2):189. Available from: /pmc/articles/PMC3546795/

55. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics [Internet]. 2013 May 1 [cited 2023 Aug 2];14(1). Available from: https://pubmed.ncbi.nlm.nih.gov/23631413/

56. van Iterson M, Tobi EW, Slieker RC, den Hollander W, Luijk R, Slagboom PE, et al. MethylAid: visual and interactive quality control of large Illumina 450k datasets. Bioinformatics [Internet]. 2014 Dec 1 [cited 2021 Jul 28];30(23):3435–7. Available from: https://academic.oup.com/bioinformatics/article/30/23/3435/207545

57. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics [Internet]. 2014 May 15 [cited 2022 Feb 25];30(10):1363–9. Available from: https://academic.oup.com/bioinformatics/article/30/10/1363/267584

58. Soerensen M, Hozakowska-Roszkowska DM, Nygaard M, Larsen MJ, Schwämmle V, Christensen K, et al. A Genome-Wide Integrative Association Study of DNA Methylation and Gene Expression Data and Later Life Cognitive Functioning in Monozygotic Twins. Front Neurosci. 2020 Apr 9;14:517712.

59. Therneau T. A Package for Survival Analysis in R. [Internet]. 2023. Available from: https://cran.r-project.org/package=survival

60. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA [Internet]. 1982 May 14 [cited 2023 Aug 2];247(18):2543–6. Available from: https://jamanetwork.com/journals/jama/fullarticle/372568

61. Van Der Weele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value.

Ann Intern Med [Internet]. 2017 Aug 15 [cited 2023 Aug 2];167(4):268–74. Available from: https://pubmed.ncbi.nlm.nih.gov/28693043/

62. Korhonen T, Hjelmborg J, Harris JR, Clemmensen S, Adami HO, Kaprio J. Cancer in twin pairs discordant for smoking: The Nordic Twin Study of Cancer. Int J Cancer [Internet]. 2022 Jul 7 [cited 2023 Aug 2];151(1):33. Available from: /pmc/articles/PMC9304125/

63. Xu Z, Xie C, Taylor JA, Niu L. ipDMR: identification of differentially methylated regions with interval P-values. Bioinformatics [Internet]. 2021 Mar 1 [cited 2023 Aug 2];37(5):711–3. Available from: https://pubmed.ncbi.nlm.nih.gov/32805005/

64. Xu Z, Niu L, Taylor JA. The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines. Clin Epigenetics [Internet]. 2021 Dec 1 [cited 2023 Aug 2];13(1). Available from: /pmc/articles/PMC8662917/

65. Jetté M, Sidney K, Blümchen G. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. Clin Cardiol [Internet]. 1990 Aug 1 [cited 2023 Aug 14];13(8):555–65. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/clc.4960130809

Figures:

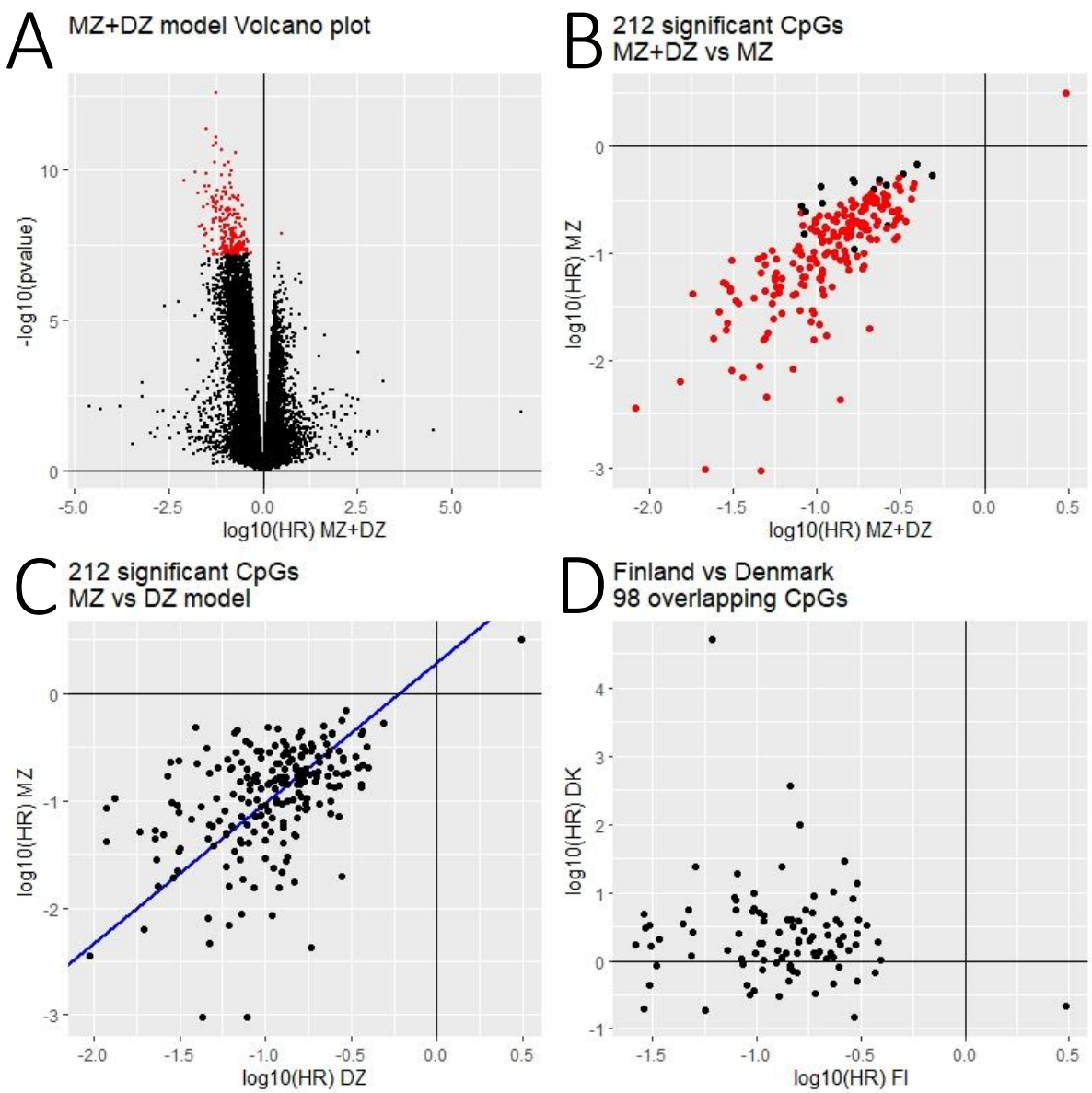

Figure 1: Results on the survival modeling for individual CpG sites associated with breast cancer; A) Volcano plot of Model 1 (MZ+DZ) with CpG sites significantly associated with breast cancer marked in red; B) Comparison between Model 1 (MZ+DZ) and Model 2 (MZ) with significant CpG sites from Model 1 validated in Model 2 marked in red; C) Comparison between Model 2 (MZ) and Model 3 (DZ) with a regression line in blue (regression coefficient = 1.31, p=0.001); D) Comparison between Model 1 (MZ+DZ, Finnish data) and Model 2R (MZ+DZ, Danish data).

Tables:

Table 1: Comparison of the MZ and DZ pairs for data relevant to the survival modeling.

	MZ, 32 pairs		DZ, 76 pairs	
Platform per pair				
450K	2		2	
EPIC	30		74	
Age at sampling in years				
Mean	57.09		55.54	
SD	10.56		9.64	
Min	38.60		38.12	
Max	79.19		78.91	
Age at study entry in years				
Mean	34.18		33.62	
SD	10.43		9.20	
Min	19.44		18.52	
Max	54.56		60.49	
Age at diagnosis/ censoring in years	Cases	Controls	Cases	Controls
Mean	66.67	75.55	66.80	75.06
SD	9.53	9.98	9.12	9.27
Min	46.79	60.07	47.53	46.60
Max	87.21	94.76	86.52	91.73
Time to diagnosis in years				
Mean	9.58	-	11.26	-
SD	6.67	-	6.61	-
Min	0.53	-	0.87	-
Max	21.40	-	23.10	-
Reason for censoring				
Breast cancer	32 (100%)	-	76 (100%)	-
Death	-	4 (13%)	-	14 (18%)
End of study	-	28 (87%)	-	62 (82%)

Table 2: Comparison of known breast cancer risk factors between MZ and DZ twin pairs.

Risk factor	MZ, 32 pairs		DZ, 76 pairs	
	Cases	Controls	Cases	Controls
Oral contraceptive use [a]				
Never	18 (67%)	19 (63%)	36 (48%)	40 (55%)
Ever	9 (33%)	11 (37%)	39 (52%)	33 (45%)
NA [c]	5	2	1	3
Age at first birth in years [b]				
Mean	26.69	26.55	25.48	25.39
SD	5.96	5.15	4.62	5.34
NA, incl. nulliparous women [c]	16	12	15	22
Number of children incl. nulliparous women [b]				
Mean	1.03	1.13	1.76	1.41
SD	1.40	1.02	1.34	1.43
NA [c]	2	1	0	0
Number of nulliparous women	14	11	15	22
BMI in kg/m^2 [a]				
Mean	21.72	21.49	22.29	22.78
SD	2.89	2.60	3.56	3.58
NA [c]	5	2	3	4
Alcohol use in gram ETOH/day [a]				
Mean	3.12	2.57	3.90	4.17
SD	3.51	2.19	3.61	4.12
NA [c]	5	2	1	3
Leisure time physical activity MET [d] hours/day [a]				
Mean	0.76	0.96	1.38	1.09
SD	0.72	1.22	2.61	1.44
NA [c]	5	2	7	10
Social class [a]				
Working on a monthly or hourly salary basis	1 (4%)	3 (10%)	3 (4%)	4 (5%)
Working contractual basis	10 (37%)	14 (47%)	30 (40%)	28 (38%)
Self-employed	6 (22%)	5 (17%)	22 (29%)	22 (30%)
Farmers	3 (11%)	2 (7%)	9 (12%)	4 (5%)
Currently not working	0	1 (4%)	3 (4%)	8 (11%)
Never working	7 (26%)	5 (17%)	8 (11%)	7 (9%)
NA [c]	5	2	1	3
Education [a]				
Primary school or less	11 (41%)	14 (47%)	45 (60%)	44 (60%)
Middle school	8 (30%)	6 (20%)	12 (16%)	14 (19%)
Highschool graduate	8 (30%)	10 (33%)	17 (23%)	15 (20%)
other	0	0	1 (1%)	0
NA [c]	5	2	1	3
Relationship status [a]				
Not in relationship (single, divorced, widowed)	15 (56%)	10 (33%)	24 (32%)	27 (37%)
In relationship (married or unmarried)	12 (44%)	20 (67%)	51 (68%)	46 (63%)
NA [c]	5	2	1	3

a: Based on 1975 questionnaire.
b: Based on data from the Finnish Population Registry.
c: Not applicable, number of individuals with missing data.
d: MET (metabolic equivalent) is a standardized measure for physical activity. One MET equal to 210 ml O2/kg/hour and represents the oxygen consumption while sitting at rest. (65)