Gene expression profiling of peripheral blood and endometrial cancer risk factors: systems epidemiology approach in the NOWAC Postgenome Cohort Study.

Oxana Gavrilyuk[1], Igor Snapkov[1,2], Jean-Christophe Thalabard[3], Lars Holden[4], Marit Holden[4], Hege M. Bøvelstad[1], Vanessa Dumeaux[5], Eiliv Lund[1,6]

Corresponding author: Oxana A. Gavrilyuk; Institutt for samfunnsmedisin, Det helsevitenskapelige fakultet, UiT Norges arktiske universitet, 9037 Tromso, Norway; phone: +47 77645126; e-mail: oxana.gavrilyuk@uit.no

Running title: Gene expression and endometrial cancer risk

Keywords: Endometrial cancer; gene expression; systems epidemiology; risk factors; prospective study;

Abbreviations:

LNYM - lifetime number of years of menstruation

BMI – body mass index

CI – confidence interval

EC – endometrial cancer

MHT - menopausal hormone therapy

NOWAC – The Norwegian Women and Cancer Study

OC – oral contraceptives

FDR – false discovery rate

GSEA - gene set enrichment analysis

Affiliations:

1. Department of Community Medicine, Faculty of Health Sciences, UiT, The Arctic University of Norway, Tromsø, Norway

2. Department of Immunology, University of Oslo and Oslo University Hospital-Rikshospitalet, Oslo, Norway
3. MAP5, MRC CNRS 8145, Université Paris Descartes, Sorbonne Paris Cite, Paris, France MAP5, MRC CNRS 8145, Université Paris Descartes, Sorbonne Paris Cite, Paris, France
4. Norwegian Computing Center, Oslo, Norway
5. PERFORM Center, Concordia University, Montreal, Canada
6. The Cancer Registry of Norway, Oslo, Norway

Funding: This study was supported by funding from the Northern Norway Regional Health Authority (Helse-Nord RHF) and by a grant from the European Research Council (ERC-AdG 232997 TICE).

Conflict of interest: The authors declare that they have no conflict of interest.

Author's contributions:

OG, EL, VD, and IS designed the study and interpreted the results. EL is a PI of the NOWAC Study. OG and IS constructed the tables and drafted the manuscript. MH and LH carried out the statistical analyses and contributed to drafting the manuscript. HMB participated in initial statistical analysis and critically revised the manuscript. VD and JCT contributed to the interpretation of the data, and the drafting and revision of the manuscript. All authors read and approved the final manuscript.

Abstract

Background:

Increasing incidence of endometrial cancer (EC), the most common gynecologic cancer in the world, requires extensive search for novel preventive tools and early intervention approaches. Several factors, including parity status, breastfeeding duration, use of oral contraceptives

(OC), coffee consumption, BMI, use of hormone replacement therapy (HRT), and lifetime number of years of menstruation have previously been reported to modify EC risk. However, establishment of reliable predictive models is impossible without knowledge on genetic changes prior to diagnosis. In this work, we aimed to establish if known EC risk factors influence peripheral blood gene expression in a prospective design.

Methods:

First, we selected variables that were shown to have an impact on EC risk in the whole Norwegian Women and Cancer (NOWAC) cohort (165 000 women). Then, we tested the association between these variables and changes in gene expression profiles in blood in a nested case-control study (79 case-control pairs) of women from the NOWAC postgenome cohort. Lastly, we undertook a gene set enrichment analysis (GSEA).

Results:

When we looked at overall gene expression, we found no difference between EC cases and controls. Introduction of parity status into the statistical model, revealed changes in expression of 1379 genes (false discovery rate (FDR) 20%) in controls, while we did not observe any expression changes in cases. 27 genes (FDR 20%) were associated with BMI increase in controls, whereas there was no association between changes in BMI and gene expression in women with EC. In GSEA, the major part of significantly enriched gene sets (2407, FDR 20%) were attributed to parity increase among cancer-free women.

Conclusions:

We found that increased number of parities and elevated BMI do not change peripheral blood gene expression in women diagnosed with EC later in life. The descriptive study design does not allow us to provide accurate explanation of our findings in biologic terms but this work brings solid background for further research on the development of predictive EC risk models.

Introduction

Endometrial cancer (EC) is s the most common and second most lethal gynecological cancer among women worlwide with 774 new cases registered in 2016 in Norway [1, 2]. While the incidence and mortality rates of several other cancers have plateaued or decreased in the last decade, the incidence of EC has been rising globally, with the highest increase found in countries that have undergone a drastic change in living standards and lifestyles [3, 4]. In particular, changes in reproductive factors (eg declines in parity) combined with increase in obesity prevalence might explain the rise in EC incidence associated with socioeconomic transition [5].

Both descriptive and analytic epidemiological approaches can help health officials appropriately target prevention and control activities, but provide only limited information on the biological processes underlying the cause-effect relationship between a risk factor and the disease. Technological advances in genomic profiling provide epidemiologists with the opportunity to integrate molecular data into etiologic studies in order to decypher the carcinogenic process.

So far, most studies have been focused on the development of tools to reclassify endometrial tumors according to their molecular features and stratify patients according to risk of metastases and recurrence [6-8]. We recently showed that an in-depth analysis of multi-tissue genomic profiles could be a crucial addition to the personalized assessment of an individual's biology and health [9].

As a major defense and transport system, blood cells can adjust expression of their genes in response to various environmental factors and pathological conditions. We previously highlighted several specific behavioral programs, such as metabolism or signaling, deregulated in the individual's blood cells that are associated with biological and/or

pathological responses to a given condition in the general population (eg smoking) [10], cancer-related risk factors (article in press), and health status (eg breast cancer diagnosis) [12, 13].

In this study, we aim to investigate the associations between known EC risk factors and gene expression changes in blood cells differential between women who will develop EC and controls. This will provide insight into biological processes underlying lifestyle/exposure EC risk factors that might explain incidence of EC.

Material and methods

The NOWAC Study

The NOWAC Study is a national population-based cohort study which include Norwegian women aged between 30-70 randomly drawn from the Norwegian Central Population Register [14]. Starting from 1991 and within 4-6 years intervals, these women filled in questionnaires with focus on lifestyle and health. Of this original cohort of about 172 000 women, more than 45 000 women born between 1943 and 1957 provided blood samples between 2003 and 2006 and filled an additional 2-page questionnaire to constitute the NOWAC Postgenome Cohort [14]. PAXgene tubes (PreAnalytiXGmbH, Hembrechtikon, Switzerland) were used to prevent RNA degradation after blood sampling and allowed genome–wide analyses of blood gene expression profiles. Through the linkage to the Cancer Registry of Norway and Register of death certificates in Statistics Norway, we have identified 88 women from the NOWAC Postgenome Cohort who developed EC between the time of blood sampling and December 31, 2008 (end of follow-up). Of these 88 individuals, four were excluded, as blood samples were not received and stored at -80C within 4 days after blood collection . To ensure the same storage time and age between cases and controls, the controls, who did not receive any cancer

diagnosis, were drawn at random from the NOWAC Postgenome Cohort but matched by time of blood collection and birth year. Matched case-control pairs of blood samples (n=84 pairs) were sent to the Genomics Core Facility at the Norwegian University of Science and Technology (NTNU) for microarray gene expression profiling in January 2011.

Assessment of covariates and calculation of lifetime number of years of menstruation

Information on the covariates age at menarche, age at menopause, number of full-term pregnancies, duration of breastfeeding, height, weight, oral contraceptive use, and smoking status was taken from NOWAC questionnaires (series from years 2002-2005). Self-reported height and weight were used to calculate BMI in kg/m^2. Parity and breastfeeding variables are generally reported to have a good validity in NOWAC Study [15]. The lifetime number of years of menstruation (LNYM) count the number of years between age at menarche and age at menopause, minus the cumulative duration of full-term pregnancies (calculated as the number of full-term pregnancies, including live and stillbirths, times 0.75 years), duration of breastfeeding (calculated as the cumulative number of months of breastfeeding in all pregnancies), and duration of OC use [16].

Laboratory procedures

In order to minimize the technical variability, each control sample was processed together with the matching case sample throughout RNA extraction, amplification and hybridization. Total RNA was isolated using the PAXgene Blood RNA Isolation Kit (Preanalytix, Qiagen, Hilden, Germany) following the manufacturer's instructions. RNA quantity and purity were assessed by the NanoDrop ND1000 spectrophotometer (Thermo Scientific, Wilmington, Delaware, USA) and Agilent bioanalyzer (Agilent Technologies, Palo Alto, CA, USA RNA amplification was performed in 96-well plates using 300 ng of total RNA and the Illumina® TotalPrep™-96 RNA Amplification Kit (Ambion Inc., Austin, TX, USA). Genome-wide RNA profiles were obtained using IlluminaHumanHT-12 chips version 4.

Outlier removal

Initial quality control made at the NTNU laboratory revealed two technical outliers (one 5S type degradation and one failed cRNA synthesis). These were removed along with their matching samples. Further, one case was later found to have two cancer diagnoses and was excluded along with its matching control. The data was carefully investigated to identify and remove technical outliers using the standard operating procedure for outlier removal described in [17]. In particular, probes related to genes in the human leukocyte antigen (HLA) systems were excluded (38 probes). These genes are known to be expressed strongly and have a high variance, which could affect multivariate analyses used in the outlier search. For individuals that were borderline outlier candidates, we excluded the ones with quality measures outside the range of the following thresholds: RIN value <7, 260/280 ratio <2, 260/230 ratio <1.7, and 50 < RNA < 500. Based on this outlier search we identified two additional individuals as technical outliers and excluded them along with their matching individual. In total, we investigated blood profiles from 79 women who developed EC and 79 age-matched controls.

Microarray data preprocessing and normalization

Microarray data preprocessing and analysis were performed using R v3.3.1 (http://cran.r-project.org), and tools from the Bioconductor project (http://www.bioconductor.org), adapted to our needs.

Expression profiles including 47 248 probes were adjusted for background noise using the negative control probes [18]. The data was further log2-transformed using a variance stabilizing technique [19] and finally normalized using quantile normalization. We retained probes present in at least 70% of the samples. If a gene was represented by more than one probe, the average expression of the probes was used as expression value for the gene. The probes were translated to genes using the lumiHumanIDMapping database [20]. In total, the

final data set included the expression for 7104 unique genes. Finally, we computed the differences in log2 gene expression levels for each case-control pair.

Statistical methods

We used linear models (Bioconductor R-package Limma [21] to evaluate the significance of gene-wise expression differences between cases and controls using the following intercept-only model:

$$Y_{g,c} = \alpha + \varepsilon_{g,c}$$

where $Y_{g,c} = Y_{g,c}^{case} - Y_{g,c}^{ctrl}$ is the difference in log2 gene expression level for case-control pair c and gene g and $\varepsilon_{g,c}$ is normally distributed with zero expectation.

Using the same appraoach, we tested whether gene expression changes between cases and controls were associated with EC risk factors (parity, LNYM, coffee, BMI, age of menopause, or OC use) using the following model:

$$Y_{gc} = \alpha + \beta_V^{ctrl} V_c^{ctrl} + \beta_V^{case} V_c^{case} + \varepsilon_{g,c}$$

where V_c^{ctrl} and V_c^{case} are the parity, coffee, LNYM, BMI, age of menopause, or OC use variable for case-control pair c in control and cases, respectively.

Similarly, we applied a gene set approach to determine changes in gene signatures between EC cases and controls associated with known risk factors. In this approach, the dependent variable is the difference in enrichment scores for case-control pair c and a defined gene set. The enrichment scores for eight collections (C1-C7 and H) of gene sets from the Molecular Signatures Database v6.1 (http://software.broadinstitute.org/gsea/index.jsp) [22] are obtained using the GSVA Bioconductor/R package [23]. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure for controlling FDR [26].

Distributions of known EC risk factors were compared between cases and matched controls using independent two-sided sample t-tests, Mann-Whitney U tests, and Chi square tests (R statistical package).

Ethics Statement

The NOWAC study was approved by the Norwegian Data Inspectorate and the Regional Ethical Committee of North Norway (REK). The study was conducted in compliance with the Declaration of Helsinki and all participants gave written informed consent. The linkages of the NOWAC database to national registries such as the Cancer Registry of Norway and registries on death and emigration was approved by the Directorate of Health. The women were informed about these linkages. Furthermore, the collection and storing of human biological material was approved by the REK in accordance with the Norwegian Biobank Act. Women were informed in the letter of introduction that the blood samples would be used for gene expression analyses.

Results

Study population

In total, blood gene expression profiles were analyzed from 79 women diagnosed with EC after blood collection and 79 women, matched by year of birth and time of blood sampling, who did not receive any cancer diagnosis within the same interval after blood collection. There was no significant difference in age, menarche onset, and number of children between the two groups (Table 1). In both group, about half of the women were premenopausal. In general, cases had later occurrence of menopause compared to controls. Additionally, we observed a difference in LNYM with trend towards increase in cases. In our study, cumulative breastfeeding duration was the lowest in controls. This can be explained by a large spread in reported time of breastfeeding by women with EC. There were more OC users among cases. Our study population was overweight, particularly, 64.6% of women with EC diagnosis had BMI >25. Controls drank slightly more coffee.

Differential blood gene expression profiles associated with EC diagnosis and risk factors in cases and controls

After preprocessing, the study dataset included expression values for 7104 genes. In overall analysis we were unable to identify any significant differences in gene expression profiles between cases and controls. The same negative result was obtained when we performed separate analysis for each year before diagnosis.

Then, we tested the hypothesis that the expression of some genes in the blood of either cases or controls might be influenced by a set of variables modulating EC risk. Among the cases, there was no relationship between any of the variables used and log gene expression. In controls, we observed no differentially expressed genes when using a model that included either of the following: coffee consumption, age of menopause, use of OC. Variations in BMI and number of pregnancies had the strongest impact on the gene expression in controls. Increasing parity was related to expression differences of 1379 genes (FDR 20%). Higher BMI altered the expression of 8, 17, and 27 genes at FDR 10%, 15%, and 20% respectively. Of note, for both BMI and parity, the major number of top 10 genes were downregulated (Table 2).

Gene set enrichment analysis

For GSEA, we used all collections available at MSigDB (Molecular Signatures Database, http://software.broadinstitute.org/gsea/index.jsp) [27]. In women diagnosed with EC, totally, we identified 3 significantly enriched gene sets (FDR 20%), of which 1 was parity-associated; 1 was enriched along with BMI increase, and 1 was linked to the use of OC (Table S1). Among the controls, GSEA revealed 2415 enriched gene sets (FDR 20%), where 2407 were attributed to parity (Figure 2). Remarkably, the biggest part of these gene sets (786 at FDR 15% and 1184 at FDR 20%) were from C7 collection (immunologic gene sets).

Discussion

In a large prospective cohort, we studied the possibility to trace blood gene expression changes prior EC diagnosis. We did not observe any significant differences in expression signatures between cancer-free controls and women with EC when compared directly. Interestingly, BMI and parity, also known to be associated with EC incidence, were associated with significant changes in blood expression profiles in controls but not in cases. To our knowledge, this is the first study demonstrating pre-diagnostic blood gene expression differences between EC cases and matched controls utilizing systems epidemiology approach.

Association between high BMI and increased risk of EC development is well established [28]. Moreover, overweight patients with EC have higher risk of death with relative risk up to 6.25 in patients with BMI>40 compared to normal weight women [29]. The main obesity-associated pathways contributing to EC development include augmented estrogen and estrogen metabolites synthesis, presence of chronic inflammation, and insulin resistance [28]. In our single gene analysis, unexpectedly, we did not observed changes in expression associated with BMI increase in EC cases. In turn, we observed a number of genes demonstrating differential expression in controls. This finding might be explained by the fact that the difference in BMI between cases and controls in the study cohort was modest (Table 1). In GSEA, we identified 3 significantly enriched gene sets from C2 collection (1 in cases and 2 in controls), and 1 gene set from H collection. Notably, "HALLMARK_TGF_BETA_SIGNALING" gene set from H collection was enriched when the BMI of controls increased. This finding is in line with other studies demonstrating the involvement of disturbed TGFβ signaling in EC development and progression [30, 31].

Not surprisingly, the major disparity on both single gene and gene sets levels in our study was connected to the number of pregnancies. There is a large body of evidence showing the negative correlation between parity and risk of EC [32-34]. However, recent meta-analysis

report nonlinear association between number of children and RR [5]. Indeed, in the entire NOWAC cohort, we found decrease in EC incidence rate in women with 1, 2 or 3 children compared to nulliparous (Figure S1). The elevated incidence rate of EC in women with 4 and more children is attributed to the low number of women with high number of pregnancies in the cohort and, therefore, limited sample size. Similar observations were reported by other studies [33]. Reduced time of estrogen exposure with increased parity is considered a major protective mechanism [35]. Additionally, shedding of the endometrium resulting in mechanical elimination of potentially premalignant cells is well described [5]. In the current study, we observed significant enrichment of a large number of immunologic gene sets (C7 collection in MSigDB) in controls with growing number of parities. Based on this finding, it is possible to assume that changes in the immune system associated with pregnancy may be yet another explanation of parity-dependent protection against EC. Moreover, this protective effect grows cumulatively with every new child. Previously, our group published similar observations on parity and BC protection in a larger cohort (article in press). Nevertheless, taking into account the complexity of gene sets data, limited sample size, and per se explorative design of this study, it is impossible to provide clear hypothesis on how the immune system changes in pregnancy contribute to EC protection. Therefore, further studies using both laboratory and epidemiologic design, which address the long-term effects of immune processes on endometrial tumorigenesis, are warranted.

In cases, only "HALLMARK_DNA_REPAIR" gene set was significantly associated with high parity (FDR 10%). Alterations in DNA repair machinery are well known to play a major role in EC carcinogenesis [36]. Thus, it is possible to hypothesize that mutations influencing DNA repair mechanisms could abrogate protective effect of parity on EC and lead to cancer development even in women who have given several births.

In current work, other factors with published evidences of involvement in EC development had relatively week impact on pre-diagnostic blood gene expression.

It has been demonstrated in NOWAC and by others that increased coffee consumption inversely associated with EC risk [9, 17]. Here we identified one significant gene set "HALLMARK_IL2_STAT5_SIGNALING" (FDR 20%) related to coffee drinking in controls. Recently, Gotthardt and co-authors demonstrated that maintenance of stable STAT5 level is necessary for tumor surveillance by NK cells [37]. STAT5 depleted NKs were shown to promote tumor development. Hence, impact of coffee compounds on STAT5 metabolism in immune cells can be an additional biologic substrate of protective functions.

In OC users among cases, we revealed significant enrichment of "REACTOME_HYALURONAN_METABOLISM" gene set (FDR 20%). Interestingly, despite the low significance level, second gene set from the top "REACTOME_HYALURONAN_UPTAKE_AND_DEGRADATION" was also related to hyaluronic acid metabolism. Elevated levels of hyaluronic acid in both tumor tissue and serum have been demonstrated to be involved in EC progression [38, 39]. Therefore, monitoring of hyaluronan in the blood of women using OC might be a valuable tool in endometrial cancer screening.

To what extent circulating blood cells can reflect processes that occur in tumors is still an open question. Being an easily accessible tissue, blood may serve as an ideal tool for disease prognosis, monitoring and assessment of the treatment. In this work, we attempted to discover gene expression changes in circulating cells that can be identified long before diagnosis of EC. It is important to emphasize that findings reported here need further investigation as most of the information available on role of different genes in tumorigenesis is based on tissue studies and, therefore, cannot be entirely extrapolated to the blood cells.

The main strengths of the study include prospective design, population representativeness of the cohort, complete information on cancer status, emigration and mortality obtained from national registries. The systems approach of testing epidemiologic data in the sub cohort using gene expression profiles reduces the probability of false positive findings.

Relatively small sample size and the lack of a common algorithm for the gene expression analysis are limitations of this work. In addition, FDR levels we accepted were higher than recommended but this can be justified by relatively small sample size and low p-values of genes and gene sets included.

In conclusion, we identified a number of differences in gene set enrichment profiles between cancer-free women and women with EC prior diagnosis in relation to known risk factors for EC. We believe that this integrated analysis may provide a promising background for developing a new multilevel prediction model of EC risk at population level. However, this approach should be further tested in a bigger sample size and in different populations.

References:

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D and Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136: E359-386.

[2] Cancer in Norway 2016. Cancer Registry of Norway. https://www.kreftregisteret.no/globalassets/cancer-in-norway/2016/cin-2106-070218.pdf.

[3] Torre LA, Siegel RL, Ward EM and Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer Epidemiol Biomarkers Prev 2016; 25: 16-27.

[4] Lortet-Tieulent J, Ferlay J, Bray F and Jemal A. International Patterns and Trends in Endometrial Cancer Incidence, 1978-2013. J Natl Cancer Inst 2017;

[5] Doepke M. Accounting for fertility decline during the transition to growth. Journal of Economic Growth 2004; 9: 347–383.

[6] O'Mara TA, Zhao M and Spurdle AB. Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. Sci Rep 2016; 6: 36677.

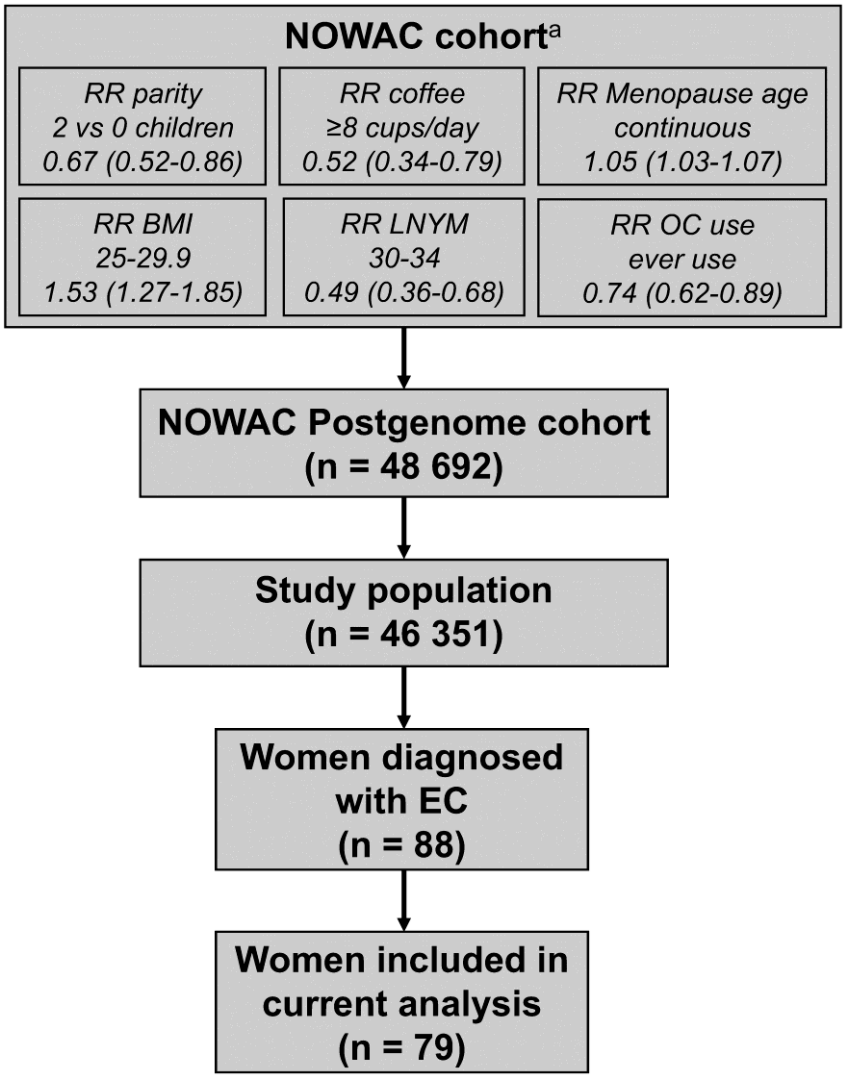
[7] Saghir FS, Rose IM, Dali AZ, Shamsuddin Z, Jamal AR and Mokhtar NM. Gene expression profiling and cancer-related pathways in type I endometrial carcinoma. Int J Gynecol Cancer 2010; 20: 724-731.

[8] Sung CO and Sohn I. The expression pattern of 19 genes predicts the histology of endometrial carcinoma. Sci Rep 2014; 4: 5174.

[9] Dumeaux V, Fjukstad B, Fjosne HE, Frantzen JO, Holmen MM, Rodegerdts E, Schlichting E, Borresen-Dale AL, Bongo LA, Lund E and Hallett M. Interactions between the tumor and the blood systemic response of breast cancer patients. PLoS Comput Biol 2017; 13: e1005680.
[10] Dumeaux V, Olsen KS, Nuel G, Paulssen RH, Borresen-Dale AL and Lund E. Deciphering normal blood gene expression variation--The NOWAC postgenome study. PLoS Genet 2010; 6: e1000873.
[11] Breast cancer paper
[12] Lund E and Dumeaux V. Systems epidemiology in cancer. Cancer Epidemiol Biomarkers Prev 2008; 17: 2954-2957.
[13] Lund E, Holden L, Bovelstad H, Plancade S, Mode N, Gunther CC, Nuel G, Thalabard JC and Holden M. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. BMC Med Res Methodol 2016; 16: 28.
[14] Lund E, Dumeaux V, Braaten T, Hjartaker A, Engeset D, Skeie G and Kumle M. Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol 2008; 37: 36-41.
[15] Lund E, Kumle M, Braaten T, Hjartaker A, Bakken K, Eggen E and Gram TI. External validity in a population-based national prospective study--the Norwegian Women and Cancer Study (NOWAC). Cancer Causes Control 2003; 14: 1001-1008.
[16] Gavrilyuk O, Braaten T, Weiderpass E, Licaj I and Lund E. Lifetime number of years of menstruation as a risk index for postmenopausal endometrial cancer in the Norwegian Women and Cancer Study. Acta Obstet Gynecol Scand 2018.
[17] Bovelstad H, Holsbo E, Bongo L and Lund E. A Standard Operating Procedure For Outlier Removal In Large-Sample Epidemiological Transcriptomics Datasets. bioRxiv 2017; https://doi.org/10.1101/144519.
[18] Shi W, Oshlack A and Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res 2010; 38: e204.
[19] Lin SM, Du P, Huber W and Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res 2008; 36: e11.
[20] Du P, Feng G, Kibbe W and Lin S. lumiHumanIDMapping: Illumina Identifier mapping for Human. R package version 1.10.1. 2016.
[21] Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology 2004; 3: 1-25.
[22] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-15550.
[23] Hänzelmann S, Castelo R and Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 2013; 14: 7.
[24] Reiner A, Yekutieli D and Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19: 368-375.
[28] Onstad MA, Schmandt RE and Lu KH. Addressing the Role of Obesity in Endometrial Cancer Risk, Prevention, and Treatment. J Clin Oncol 2016; 34: 4225-4230.
[29] Calle EE, Rodriguez C, Walker-Thurmond K and Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med 2003; 348: 1625-1638.
[30] Li Q. Transforming growth factor beta signaling in uterine development and function. J Anim Sci Biotechnol 2014; 5: 52.
[31] Piestrzeniewicz-Ulanska D, Brys M, Semczuk A, Rechberger T, Jakowicki JA and Krajewska WM. TGF-beta signaling is disrupted in endometrioid-type endometrial carcinomas. Gynecol Oncol 2004; 95: 173-180.
[32] Kvale G, Heuch I and Ursin G. Reproductive factors and risk of cancer of the uterine corpus: a prospective study. Cancer Res 1988; 48: 6217-6221.
[33] Dossus L, Allen N, Kaaks R, Bakken K, Lund E, Tjonneland A, Olsen A, Overvad K, Clavel-Chapelon F, Fournier A, Chabbert-Buffet N, Boeing H, Schutze M, Trichopoulou A, Trichopoulos D, Lagiou P, Palli D, Krogh V, Tumino R, Vineis P, Mattiello A, Bueno-de-Mesquita HB, Onland-Moret NC, Peeters PH, Dumeaux V, Redondo ML, Duell E, Sanchez-Cantalejo E, Arriola L, Chirlaque MD, Ardanaz E, Manjer J, Borgquist S, Lukanova A, Lundin E, Khaw KT, Wareham N, Key T, Chajes V, Rinaldi S, Slimani N, Mouw T, Gallo V and Riboli E. Reproductive risk factors and endometrial cancer: the European Prospective Investigation into Cancer and Nutrition. Int J Cancer 2010; 127: 442-451.
[34] Schonfeld SJ, Hartge P, Pfeiffer RM, Freedman DM, Greenlee RT, Linet MS, Park Y, Schairer C, Visvanathan K and Lacey JV, Jr. An aggregated analysis of hormonal factors and endometrial cancer risk by parity. Cancer 2013; 119: 1393-1401.

[35] Chen Q, Tong M, Guo F, Lau S and Zhao M. Parity Correlates with the Timing of Developing Endometrial Cancer, But Not Subtype of Endometrial Cancer. J Cancer 2015; 6: 1087-1092.
[36] Masuda K, Banno K, Yanokura M, Kobayashi Y, Kisu I, Ueki A, Ono A, Asahara N, Nomura H, Hirasawa A, Susumu N and Aoki D. Relationship between DNA Mismatch Repair Deficiency and Endometrial Cancer. Mol Biol Int 2011; 2011: 256063.
[37] Gotthardt D, Putz EM, Grundschober E, Prchal-Murphy M, Straka E, Kudweis P, Heller G, Bago-Horvath Z, Witalisz-Siepracka A, Cumaraswamy AA, Gunning PT, Strobl B, Muller M, Moriggl R, Stockmann C and Sexl V. STAT5 Is a Key Regulator in NK Cells and Acts as a Molecular Switch from Tumor Surveillance to Tumor Promotion. Cancer Discov 2016; 6: 414-429.
[38] Paiva P, Van Damme MP, Tellbach M, Jones RL, Jobling T and Salamonsen LA. Expression patterns of hyaluronan, hyaluronan synthases and hyaluronidases indicate a role for hyaluronan in the progression of endometrial cancer. Gynecol Oncol 2005; 98: 193-202.
[39] Yabushita H, Kishida T, Fusano K, Kanyama K, Zhuo L, Itano N, Kimata K and Noguchi M. Role of hyaluronan and hyaluronan synthase in endometrial cancer. Oncol Rep 2005; 13: 1101-1105.

Figure 1. Study population.



Notes: a. Relative risk estimates of EC in the Norwegian Women and Cancer Study calculated using multivariable model adjusted for BMI, use of HRT, use of OC, smoking and alcohol consumtion

Figure 2. Number of significantly enriched gene sets in association with increasing parity.

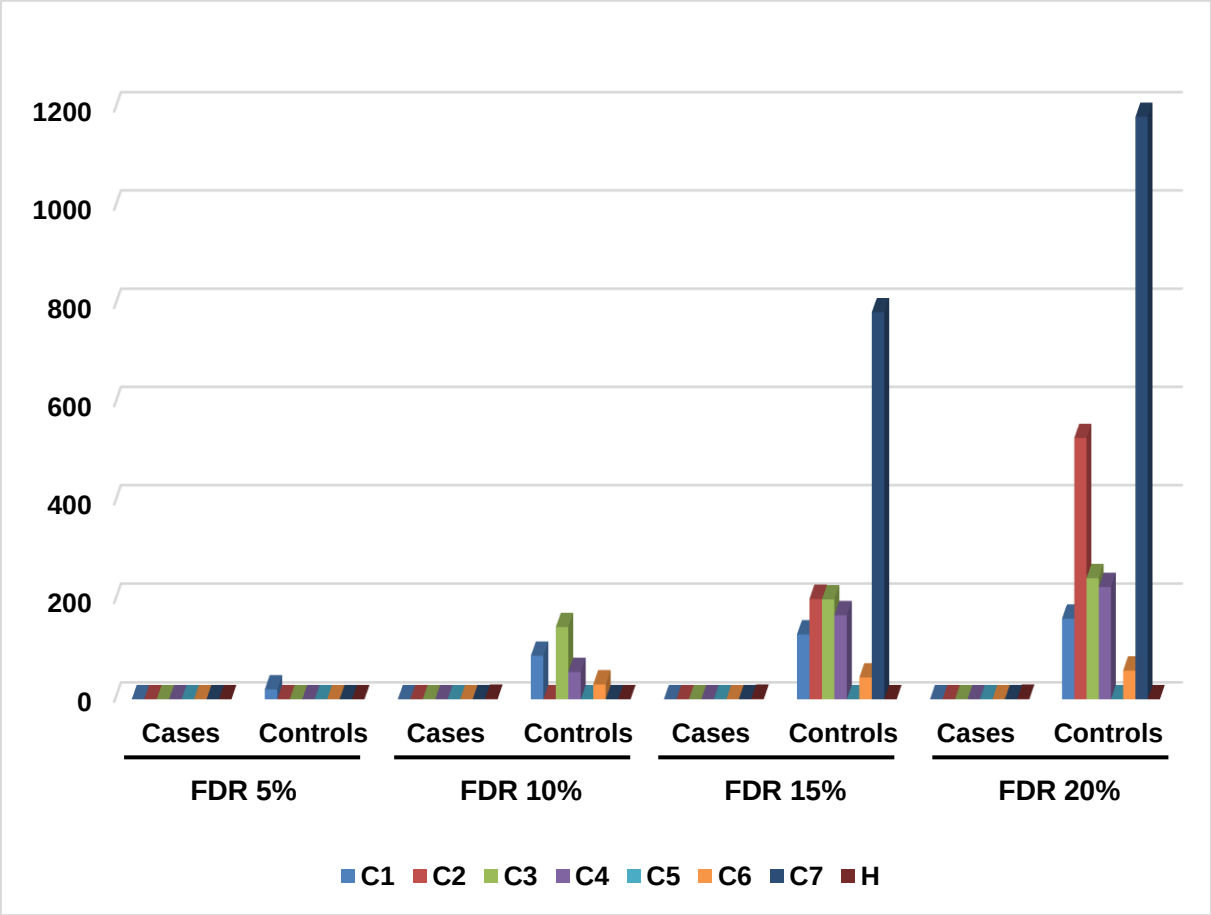


Table 1. Baseline characteristics of the study (N = 79 case/control pairs)

Characteristics	**EC cases**	**Controls**
Age (mean, ±SE)	49.3 (6.3)	49.6 (6.5)
Menopausal status *Premenopausal* *Postmenopausal*	 38 (56 %) 30 (44 %)	 41 (46 %) 49 (54 %)
Age at menopause (mean, ±SE)	49.9 (4.2)	47.5 (5.5)
Age at menarche (mean, ±SE)	12.9 (1.4)	13.2 (1.3)
Parity 0 1 2 3 4 5	 8 11 38 17 5 0	 8 9 35 17 9 1
Cumulative duration of breastfeeding (mean, ±SE)	13.5 (17.0)	10.4 (7.2)
Ever consumption of oral contraceptives (%)	43%	30.7%
LNYM <25 25-29 30-34 35-39 40+	 10.7% 10.7% 36.0% 34.7% 8.0%	 15.4% 14.1% 41.0% 26.9% 2.6%
Body mass index (mean, ±SE) < 25 (%) > 25 (%)	27.5(5.4) 35.4% 64.6%	25.8 (4.9) 58.2% 41.8%
Ever coffee consumption (%)	92.4%	93.7%

Table 2. Top 10 differentially expressed genes associated with parity and BMI.

Parity				
Gene	**logFC[a]**	**p-value**	**q-value[b]**	**Function**
NUDT22	-0.081	8.15E-05	0.163	UDP-glucose and UDP-galactose hydrolase
SH2B2	-0.104	8.34E-05	0.163	Regulator of tyrosine kinase receptor activity
NUP188	-0.056	0.000123829	0.163	Nuclear pore complex involved in the flow of various substances between the cytoplasm and nucleoplasm
TRIP12	0.134	0.000157622	0.163	E3 ubiquitin-protein ligase, involved in regulation of DNA repair
APBA3	-0.094	0.000204851	0.163	Involved in signal transduction and synaptic transmission
CYB5A	0.055	0.000221835	0.163	Electron carrier, regulates hemoglobin metabolism
CEP250	-0.053	0.000252363	0.163	Required for interphase progression of the cell cycle
NRM	-0.073	0.000258652	0.163	Encodes protein residing within the inner nuclear membrane. May be involved in apoptosis
PLRG1	0.097	0.000348695	0.163	Regulator of alternative spicing
TWF2	-0.083	0.000479402	0.163	involved in motile processes and endocytosis regulation
BMI				
Gene	**logFC[a]**	**p-value**	**q-value[b]**	**Function**
ALS2	0.016	2.69E-05	0.099	GTPase regulator. Involved in the development of spinal neurons
TAOK1	0.026	3.11E-05	0.099	Involved in p38 MAPK signaling, apoptosis regulation and cytoskeleton maintenance
ZZEF1	-0.019	7.34E-05	0.099	Involved in calcium ion binding.
DNAJB1	-0.020	7.35E-05	0.099	Stimulates ATP hydrolysis and promotes folding and unfolding of proteins
PROSC	-0.017	7.87E-05	0.099	Involved in homeostasis regulation of pyridoxal 5-phosphate (active form of B6 vitamin)
H2AFY	-0.023	8.35E-05	0.099	Histone-coding gene that represses transcription and inactivates X chromosome

EDEM1	-0.021	0.000109136	0.099	Involved in protein processic in endoplasmic reticulum.
ZBTB44	0.024	0.000111261	0.099	Zinc finger protein realted to nucleic acid binding
SFRS9	-0.019	0.000132804	0.105	Regulates mRNA maturation.
ANKRD11	-0.020	0.000156181	0.110	Inhibits ligand-dependent activation of transcription

Notes: [a]LogFC is the estimated log-fold change in gene expression when the parity increases continuously. [b]q-value is an FDR adjusted p-value.

1

Supplementary Table 1. Gene set enrichment analysis.

	Number of significant gene sets (FDR 5% – 10% – 15% – 20%)	
Parity	**Cases**	**Controls**
C1	0 - 0 - 0 - 0	20 - 88 - 131 - 163
C2	0 - 0 - 0 - 0	0 - 0 - 203 - 530
C3	0 - 0 - 0 - 0	0 - 146 - 202 - 245
C4	0 - 0 - 0 - 0	0 - 55 - 170 - 227
C5	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C6	0 - 0 - 0 - 0	0 - 30 - 44 - 58
C7	0 - 0 - 0 - 0	0 - 0 - 786 - 1184
H	0 - 1 - 1 - 1	0 - 0 - 0 - 0
Coffee		
C1	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C2	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C3	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C4	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C5	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C6	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C7	0 - 0 - 0 - 0	0 - 0 - 0 - 0
H	0 - 0 - 0 - 0	0 - 0 - 0 - 1
BMI		
C1	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C2	1 - 1 - 1 - 1	0 - 0 - 2 - 2
C3	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C4	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C5	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C6	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C7	0 - 0 - 0 - 0	0 - 0 - 0 - 0
H	0 - 0 - 0 - 0	1 - 1 - 1 - 1
LNYM		

C1	0 - 0 - 0 - 0	0 - 1 - 1 - 2
C2	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C3	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C4	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C5	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C6	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C7	0 - 0 - 0 - 0	0 - 0 - 0 - 0
H	0 - 0 - 0 - 0	0 - 0 - 1 - 1
Age of menopause		
C1	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C2	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C3	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C4	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C5	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C6	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C7	0 - 0 - 0 - 0	0 - 0 - 0 - 0
H	0 - 0 - 0 - 0	0 - 0 - 0 - 0
OC use		
C1	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C2	0 - 0 - 0 - 1	0 - 0 - 0 - 0
C3	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C4	0 - 0 - 0 - 0	1 - 1 - 1 - 1
C5	0 - 0 - 0 - 0	0 - 0 - 1 - 1
C6	0 - 0 - 0 - 0	0 - 0 - 0 - 0
C7	0 - 0 - 0 - 0	0 - 0 - 0 - 0
H	0 - 0 - 0 - 0	0 - 0 - 0 - 0

1

Supplementary Figure 1. EC risk according to parity