Research Article

Characterization of a Genomic Signature of Pregnancy Identified in the Breast

Ilana Belitskaya-Lévy[1], Anne Zeleniuch-Jacquotte[2], Jose Russo[4], Irma H. Russo[4], Pal Bordás[6], Janet Åhman[6], Yelena Afanasyeva[2], Robert Johansson[7], Per Lenner[7], Xiaochun Li[1], Ricardo López de Cicco[4], Suraj Peri[5], Eric Ross[5], Patricia A. Russo[4], Julia Santucci-Pereira[4], Fathima S. Sheriff[4], Michael Slifker[5], Göran Hallmans[8], Paolo Toniolo[2,3,9], and Alan A. Arslan[2,3]

Abstract

The objective of this study was to comprehensively compare the genomic profiles in the breast of parous and nulliparous postmenopausal women to identify genes that permanently change their expression following pregnancy.

The study was designed as a two-phase approach. In the discovery phase, we compared breast genomic profiles of 37 parous with 18 nulliparous postmenopausal women. In the validation phase, confirmation of the genomic patterns observed in the discovery phase was sought in an independent set of 30 parous and 22 nulliparous postmenopausal women. RNA was hybridized to Affymetrix HG_U133 Plus 2.0 oligonucleotide arrays containing probes to 54,675 transcripts, scanned and the images analyzed using Affymetrix GCOS software. Surrogate variable analysis, logistic regression, and significance analysis of microarrays were used to identify statistically significant differences in expression of genes. The false discovery rate (FDR) approach was used to control for multiple comparisons. We found that 208 genes (305 probe sets) were differentially expressed between parous and nulliparous women in both discovery and validation phases of the study at an FDR of 10% and with at least a 1.25-fold change. These genes are involved in regulation of transcription, centrosome organization, RNA splicing, cell-cycle control, adhesion, and differentiation. The results provide initial evidence that full-term pregnancy induces long-term genomic changes in the breast. The genomic signature of pregnancy could be used as an intermediate marker to assess potential chemopreventive interventions with hormones mimicking the effects of pregnancy for prevention of breast cancer. *Cancer Prev Res; 4(9); 1457–64. ©2011 AACR.*

Introduction

It is well established that a pregnancy completed to term at a young age reduces the risk of breast cancer later in life. This long-term reduction in risk has been attributed to the early differentiation of breast tissue, which otherwise remains undifferentiated and susceptible to carcinogenic insults (1, 2). Because a first full-term pregnancy (FTP) and ensuing breastfeeding are the most significant physiologic events which transform the breast from an immature to a fully mature organ, Russo and colleagues hypothesized that having completed at least 1 FTP would result in a specific, detectable genomic signature in the breast (3–6). Only 1 study to date has examined gene expression in the healthy breast and it was limited to 64 genes (7).

Identification of a specific genomic fingerprint of pregnancy would open up a broad set of opportunities for understanding, and possibly preventing, breast cancer. We, therefore, undertook to compare the gene expression profiles in breast biopsy specimens of healthy parous and nulliparous volunteers from the general population, using a genome-wide approach. Because we were interested in long-term genomic changes associated with FTP, the study was focused on postmenopausal women.

Authors' Affiliations: Divisions of [1]Biostatistics and [2]Epidemiology, Department of Environmental Medicine; [3]Department of Obstetrics and Gynecology, New York University School of Medicine, New York, New York; [4]Breast Cancer Research Laboratory; [5]Department of Biostatistics and Bioinformatics, Fox Chase Cancer Center, Philadelphia, Pennsylvania; [6]Department of Radiology, Norrbotten Mammography Screening Program, Sunderby Hospital, Luleå; Departments of [7]Radiation Sciences and Oncology and [8]Public Health and Clinical Medicine, Umeå University, Umeå, Sweden; and [9]Institute of Social and Preventive Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

Materials and Methods

Study design

The study was designed to include 2 phases, a discovery phase (DP) and a validation phase (VP) with a total target sample size of 120 women (40 parous and 20 nulliparous in each phase). Recruitment was conducted without interruption between the 2 phases of the study, using the same source population. The parity distribution was reviewed after every group of 10 eligible volunteers. If the nulliparous to parous ratio differed from 1:2, recruitment was limited to the underrepresented group (usually nulliparous) until the ratio reached the 1:2 target.

Reproducibility study

To assess within- and between-laboratory reproducibility of gene expression profiles in replicate experiments, we conducted a substudy prior to the start of the discovery and validation phases. Breast tissue samples from 4 subjects were processed, and their gene expression profiles were analyzed at 3 independent laboratories (Fox Chase Cancer Center Breast Cancer Research Laboratory of Dr. Jose Russo; University of Memphis Genomic Laboratory of Dr. Thomas Sutter; and Fox Chase Cancer Center Genomic Laboratory) using identical procedures. Data were preprocessed for each of the 3 laboratories separately using the methods described in the statistical methods section. Supplementary Table S1 shows concordance correlation coefficients (8) and Pearson's correlation coefficients for within- and between-laboratory comparisons. The within-laboratory correlations were very high (Pearson's and concordance correlation coefficients: >97%) and very similar to those reported by others (9). The between-laboratory correlations were also high (Pearson's and concordance correlation coefficients: >92%).

Study population and eligibility criteria

Study subjects were recruited at the Sunderby Hospital in Luleå, Sweden, among women who have had a normal mammogram within the year prior to enrollment. Postmenopausal women (defined as lack of menstrual periods for the previous 12 months) between the ages of 50 and 69 years were approached by a research nurse who explained the study procedures and provided informed consent forms. Volunteers who signed informed consent to participate in the study and to donate biological samples for research were scheduled for an interview.

Women who reported a history of any cancer, the use of any hormonal medications in the 6 months preceding their visit, prior breast biopsy or breast implants were excluded. In the discovery phase, it became apparent that women with fatty breast (transparent mammograms) had to be excluded because of low RNA yield. As a consequence, women with similarly fatty breast were considered ineligible for biopsy in the validation phase of the study. The project was approved by the Regional Ethical Review Board for Northern Sweden at the University of Umeå, Sweden.

Data and sample collection

The study nurse obtained anthropometrical measures (height and weight) and administered the study questionnaire to eligible and consented women. The collected data included a detailed reproductive history, medical history, first-degree family history of breast cancer, smoking, and use of oral contraceptives (OC), hormone replacement therapy (HRT), and other medications.

An experienced intervention radiologist carried out all breast biopsies with a Bard Monopty (C. R. BARD Inc.) automated core biopsy instrument (14 Gauge, 10-cm long, 22-mm penetration depth) through a single small skin incision after the puncture site had been sterilized and anaesthetized (Xylocain and Adrenalin solution, 10 mg/mL + 5 μg/mL, Astrazeneca). Several (3–5) random biopsies were taken from the upper outer quadrant of 1 breast. One biopsy specimen was placed in 70% ethanol for histopathologic analysis, and the remaining ones were immediately placed in RNAlater (Ambion) solution.

The study pathologist has reviewed all tissues to make sure that research biopsies were free of atypia or cancer using criteria published previously (10). This review resulted in the exclusion of 1 study subject (see, Supplementary Fig. S1).

Sample and data blinding

Prior to sending the samples and data to laboratory at the Fox Chase Cancer Center, Philadelphia, all samples were stripped of any personal identifiers and assigned random numbers. The link between the subject's random number and subject's identifiable information was accessible only by the authorized personnel in Sweden. The laboratory personnel at the Fox Chase Cancer Center were blinded to samples' parity status and other personal information.

RNA isolation

Total RNA from the core biopsy samples was isolated using the Qiagen Allprep RNA/DNA Mini Kit, according to the manufacturer's instructions (Qiagen). Total RNA was eluted in a final volume of 60 μL (H$_2$O) and stored at −80°C until further processing. RNA quantity and quality were assessed by means of the Agilent 2100 Bioanalyzer (Agilent Technologies). The amount of total RNA yielded from the core biopsies ranged from 150 ng to 4 μg depending on the ratio of stroma to epithelial tissue.

Affymetrix microarray gene analysis

The GeneChip Expression 3′-Amplification Two-Cycle cDNA Synthesis Kit was used (Affymetrix). Double-stranded cDNA was synthesized from 100 ng of total RNA. An *in vitro* transcription reaction was then done to produce biotin-labeled cRNA from the cDNA. The cRNA was fragmented before hybridization. A hybridization cocktail, which included the fragmented target, was prepared. The hybridization cocktail was then hybridized to Affymetrix HG_U133 Plus 2.0 oligonucleotide arrays containing probes to 54,675 transcripts. Standard Affymetrix quality control measures (average background, scale

factors, and percent present calls) were applied to assess the quality of RNA samples and their subsequent labeling and hybridization, and chips that did not pass the quality control criteria were rejected. Additionally, graphical criteria based on probe-level model (PLM) analysis were applied.

Statistical Methods

Data preprocessing

Raw data from array scans were preprocessed and analyzed using the R language for statistical computing (11) and Bioconductor (12), an open source software for bioinformatics. The data were preprocessed using the Robust Multichip Average method implemented in the Bioconductor package *affy* that includes background correction, quantile normalization, and summarization of expression values (13–15). Probes for which the proportion of present calls was less than 75%, and the difference in the proportion of present calls between parous and nulliparous women was less than 25% were filtered out. Probes with low coefficient of variation across samples (below first quartile) were also removed. These filtering criteria left 19,028 probes for analysis in the discovery phase and 17,750 probes in the validation phase. The overlap between the 2 sets of probes consisted of 16,002 probes.

Batch adjustment

The microarray experiments in both phases were conducted in 8 batches. To account for potential between-batch variability, an Empirical Bayes method, implemented in the COMBAT software, developed by (16) and written in R, was used. We also corrected for batch effects in the analysis. Additionally, the quality control duplicate samples were used to evaluate the batch effects and the effectiveness of batch adjustments.

Differential gene expression

To identify genes differentially expressed between parous and nulliparous samples, we used the following 3 methods: significance analysis of microarrays (SAM; method 1; ref. 17) implemented in the R package *samr*, surrogate variable analysis (SVA; method 2; ref. 18, 19) implemented in the R package *sva*, and logistic regression analysis (LRA; method 3).

It has been shown that genetic, environmental, demographic, and technical factors may have substantial effects on gene expression (18–21). In addition to measured variables of interest, there might be sources of signal because of unknown or unmeasured factors. Leek and Storey (18) showed that failing to incorporate these sources of heterogeneity into analysis can result in both spurious and masked associations. They introduced SVA to overcome the problems caused by heterogeneity in gene expression studies and showed that SVA increases the biological accuracy and reproducibility of gene expression studies. SVA uses a residual expression matrix, obtained by removing the effects of the outcome variable (parity status in our study) on expression, to estimate, via singular value decomposition of the residual matrix, the signatures of expression heterogeneity in terms of an orthogonal basis of singular vectors. Statistical procedures are then used to assess the significance of these signatures, to identify the subset of genes driving each signature, and to form surrogate variables (SV) based on the signatures of the corresponding subsets of genes in the original expression data. The resulting SVs are used to adjust the analysis of the associations between genes and parity status. For each gene, an unadjusted P value measuring the significance of that gene as an independent predictor of the outcome variable is calculated using logistic regression that adjusts for SVs (method 2).

We also used LRA (method 3) to identify differentially expressed genes while controlling for the effects of potentially confounding factors that were measured in the study, such as body mass index (BMI), OC use history, HRT, and smoking history. To select a subset of the measured characteristics for inclusion in the LRA for adjustment, we compared them to the significant SVs derived in method 2 using the Spearman's correlation. The top 5 characteristics that were most significantly associated with 1 of the SVs in the discovery phase were selected to be adjusted for in the LRA. These characteristics were BMI, HRT duration, breast density, smoking duration, and OC use history. Other combinations of characteristics significantly associated with SVs were also adjusted for in the logistic regression analyses of parity status. For each gene, an unadjusted P value measuring the significance of that gene as an independent predictor of parity status was calculated using logistic regression adjusting for the selected variables.

The false discovery rate (FDR) approach was used to control for multiple comparisons (17, 22, 23). SAM (method 1) computes a 2-sample t test-like statistic for each gene and uses a permutation procedure to estimate FDRs which are used to select differentially expressed genes. For SVA and logistic regression, the QVALUE method (22) implemented in the R package *qvalue* was used to adjust P values for multiple comparisons. Genes with an FDR of less than 10% and at least a 1.25-fold change between parous and nulliparous samples were considered statistically significant.

Genomic signature of pregnancy in the breast

To derive a genomic predictor of pregnancy in the breast, we used 5 classification methods: "neareast shrunken centroids" method implemented in the Bioconductor package *pamr* (24), support vector machine implemented in the R library *e1071* (25), Classification and Regression Trees implemented in the R library *rpart* (26), boosted classification trees using the AdaBoost algorithm (27) and random forest implemented in the R package *randomForest* (28). The approximately 500 most significant genes based on a combined FDR and fold-change criterion were used for these analyses except for methods that carry out automatic variable selection (e.g., nearest shrunken centroids classifier). We then used the genomic classifiers identified in the discovery phase to estimate the probability of being parous

(parity score) for each woman in the validation phase. The significance of the genomic signature of parity was evaluated using the logistic regression model with FTP as the response variable and the parity score (genomic predictor) along with potential confounding variables, such as BMI, breast density, HRT, and OC duration and smoking history, as independent predictors. Sensitivity and specificity of each genomic classifier derived in the discovery phase were evaluated in the validation phase. The statisticians were blinded to the parity status of the women in the validation phase.

Results

Supplementary Fig. S1 provides a workflow of subject accrual and sample processing. A total of 389 women were interviewed between September 2008 and May 2009. Among these, 134 (34%) were excluded based on the eligibility criteria and 4 (1%) cancelled their interview. This resulted in 251 women (111 nulliparous and 140 parous) included in the study. Two (0.5%) women were excluded later in the study when it was found that one of them had breast cancer and the other one had premenopausal FSH levels. Additionally, 123 women were excluded after RNA extraction because of RNA degradation, absence of epithelial structures or an insufficient amount of RNA. The remaining 126 women (44 nulliparous and 82 parous) were included in the current study. Nineteen microarray chips were rejected based on standard Affymetrix quality control measures (average background, scale factors, and percent present calls) and based on PLM analysis. This left 107 chips for the differential expression analysis of parous versus nulliparous women: 55 in the discovery phase (37 parous and 18 nulliparous) and 52 in the validation phase (30 parous and 22 nulliparous).

Table 1 presents characteristics of parous and nulliparous women in the discovery and validation phases. There were no statistically significant differences between parous and nulliparous women within each study phase or between women in the discovery and validation phases within each parity group.

Differential gene expression

Using SVA, we identified 11 significant SVs in the discovery and 9 in the validation phase. These variables potentially have a significant effect on gene expression. Two of the SVs in both phases accounted for over 10% of the variation in gene expression (data not shown). Supplementary Tables S2A and SB show Spearman's correlation coefficients of the SVs, including batch. Prior to batch adjustment, the batch variable was significantly associated with SV1 ($\rho = 0.56$, $P = 0.01$), indicating that batch adjustment was required. It was no longer significantly associated with any of the SVs after adjustment using the COMBAT method. BMI, HRT duration, OC duration, breast density, and smoking history were significantly associated with the SVs indicating that these factors might impact gene expression. These variables were controlled for in LRA (method 3). Some of the SVs were found to be significantly associated with FTP (e.g., SV1

and SV4 in the discovery phase and SV2, SV4, and SV6 in the validation phase) and were, therefore, excluded from SVA analysis (method 2) because our objective was to identify genes significantly associated with FTP.

Table 2 presents the numbers of statistically significant genes identified using the 3 statistical methods in the discovery and validation phases at an FDR of 10% and with at least a 1.25-fold change. The numbers of differentially expressed genes were much higher in the discovery than in the validation phase. In both phases combined, depending on the statistical method used, between 228 and 288 genes were identified as differentially expressed, whereas 218 genes were identified as differentially expressed by all 3 methods. Upregulated genes were found to be more reproducible than the downregulated genes with 62% to 64% of the significant genes identified in the validation phase being also significant in the discovery phase, compared with 9% to 13% of downregulated genes. SAM and LRA yielded the highest proportion of reproducible genes: 45% of genes significant in the validation phase were also significant in the discovery phase. Using LRA, 305 probe sets identified as significantly differentially expressed in the discovery phase and confirmed in the validation phase. These genes are reported in Supplementary Table S3.

Genomic signature of pregnancy in the breast

Five genomic predictors of parity were derived in the discovery phase using the 5 classification methods described in the Statistical Methods section, and their significance was evaluated in the validation phase using logistic regression models that adjusted for clinical variables. Table 3 shows the estimated coefficients, standard errors, and P values of the genomic predictors and clinical variables in the logistic regression models applied to the validation phase with FTP as dependent variable. The results indicate that the genomic predictors derived using the discovery phase data remained significant predictors of parity in the validation phase with and without adjustment for other variables (data without adjustment are not shown). Boosted classification trees and nearest shrunken centroids consistently carried out better than the other classification methods. The prediction accuracy of the best classifiers was estimated in the validation phase to be between 65% and 75%.

Discussion

To our knowledge, this is the first study that seeks to comprehensively characterize in an unselected, population-based group of healthy volunteers the differences in gene expression in breast core biopsy specimens between parous and nulliparous women. Using a discovery validation approach, 274 upregulated and 31 downregulated probe sets were identified, which may constitute the core genomic signature that distinguishes the breast of parous postmenopausal women from that of nulliparous ones. Using supervised learning methods, we derived a genomic signature of parity in breast specimens from the subjects

Table 1. Characteristics of subjects in discovery and validation phases

	Discovery cohort			Validation cohort			Discovery vs. validation P for difference[a]	
	Parous (n = 37) Mean (SD) or N [%]	Nulliparous (n = 18) Mean (SD) or N [%]	P for difference[a]	Parous (n = 30) Mean (SD) or N [%]	Nulliparous (n = 22) Mean (SD) or N [%]	P for difference[a]	Parous	Nulliparous
Age at visit, y	60.1 (4.7)	60.8 (4.8)	0.57	58.9 (5.2)	59.5 (5.8)	0.79	0.41	0.40
Age at menarche, y	13.2 (2.0)	13.1 (1.6)	0.80	13.2 (1.6)	13.1 (1.4)	0.87	0.54	0.61
Age at menopause, y	47.5 (7.6)	49.2 (7.3)	0.46	48.7 (4.5)	49.0 (4.8)	0.95	0.80	0.61
BMI, kg/m^2	25.2 (4.1)	27.5 (6.3)	0.22	24.1 (2.4)	24.4 (3.3)	0.90	0.36	0.13
OC start age, y	22.4 (5.8)	19.5 (4.2)	0.07	22.0 (6.5)	20.2 (3.3)	0.75	0.60	0.43
OC duration, y	6.0 (6.9)	5.7 (5.3)	0.75	6.4 (7.2)	10.1 (7.2)	0.12	0.71	0.10
HRT age, y	49.3 (4.8)	45.6 (9.6)	0.60	49.1 (6.5)	49.2 (4.1)	0.98	0.87	0.64
HRT duration, y	5.75 (4.5)	10.4 (7.1)	0.10	7.2 (5.0)	6.5 (4.5) 8.0 (4.0)	0.71	0.34	0.75
OC use: no yes	5 [13%] 32 [87%]	3 [17%] 15 [83%]	1.00	7 [23%] 23 [77%]	10 [45%] 12 [55%]	0.14	0.35	0.09
HRT use: no yes	16 [43%] 21 [57%]	10 [56%] 8 [44%]	0.57	12 [40%] 18 [60%]	9 [41%] 13 [59%]	1.00	0.81	0.53
Hysterectomy: no yes	29 [78%] 8 [22%]	13 [72%] 5 [28%]	0.74	26 [87%] 4 [13%]	18 [82%] 4 [18%]	0.71	0.53	0.71
Smoking: never former current	11 [30%] 17 [46%] 9 [24%]	5 [28%] 11 [61%] 2 [11%]	0.48	14 [47%] 11 [37%] 5 [16%]	9 [41%] 11 [50%] 2 [9%]	0.59	0.35	0.73
Family history of breast cancer no yes	31 [84%] 6 [16%]	18 [100%] 0 [0%]	0.16	28 [93%] 2 [7%]	19 [86%] 3 [14%]	0.64	0.28	0.24
Pregnancies, n 1 2 >2	6 [16%] 5 [14%] 26 [70%]	–	–	3 [10%] 10 [33%] 17 [57%]	–	–	0.16	–
FTP, n 0 1 2 >2	0 9 [24%] 15 [41%] 13 [35%]	–	–	0 3 [10%] 15 [50%] 12 [40%]	–	–	0.34	–
Age at first pregnancy, mean (SD) categories, n (%) <25 25–29 ≥30	23.0 (3.9) 23 (62%) 11 (30%) 3 (8%)	–	–	23.4 (4.6) 18 (60%) 8 (27%) 4 (13%)	–	–	1.00	–
Age at first FTP, mean (SD)	24.1 (4.5)	–	–	23.9 (4.8)	–	–	0.82	–
Breast feeding, %	100%	–	–	100%	–	–	1.00	–
Breast feeding mean (SD) duration, weeks	52.2 (50.0)	–	–	63.2 (72.3)	–	–	0.88	–

[a]Based on Wilcoxon Rank Sum test for continuous variables and Fisher's exact test for categorical variables.

Table 2. Significant genes identified by the 3 methods (SAM, SVA, and LRA) using FDR equal to 10% and fold change greater than 1.25

		Methods		
		Method 1: SAM	**Method 2: SVA**	**Method 3: LRA**
Upregulated genes				
Significant genes	DP	1749	1859	1773
	VP	463	370	428
Intersection		288	228	274
Consistency[a]	DP VP	1314 (75%) 446 (96%)	1366 (73%) 351 (95%)	1344 (76%) 411 (96%)
Downregulated genes				
Significant genes	DP	1061	1058	1030
	VP	246	288	243
Intersection		30	26	31
Consistency[a]	DP VP	847 (80%) 152 (62%)	844 (80%) 146 (51%)	837 (81%) 154 (63%)
Overall				
Significant genes	DP	2810	2917	2803
	VP	709	658	671
Intersection		318	254	305
Consistency[a]	DP VP	2158 (77%) 597 (85%)	2210 (76%) 497 (76%)	2181 (78%) 565 (84%)

[a]Consistency is defined as the proportion of significantly upregulated (or downregulated) genes in the DP (VP) phase that are also upregulated (or downregulated) in the VP (DP) phase.

included in the discovery phase of the project and established that it was a significant independent predictor of parity in the subjects in the validation phase.

The genes differentially expressed in parous and nulliparous postmenopausal women are presented in Supplementary Table S3 and involved in regulation of transcription, centrosome organization, RNA splicing, cell-cycle control, adhesion, and differentiation. Among upregulated genes, *EZH2* is a member of polycomb group of proteins involved in maintaining transcriptional repression of genes. This gene acts as a tumor suppressor and also functions as a histone methyltransferase (29). *NINL*, *TRAF5*, and *SFI1* are upregulated in parous breast and involved in centrosome organization and maintaining the microtubule cytoskeleton. *CDK3*, *MCTS1*, and *SYCP2* are involved in cell-cycle control. *PRPF39*, *LUC7L3*, *HNRNPA1*, *HNRNPA2B1*, *PNN*, *PABPN1*, *RBMX*, *SNRNP200*, *PRPF4B*, and *SFPQ* are involved in RNA splicing. Among the downregulated genes,

Table 3. Significance of the genomic predictors of parity derived in the discovery phase and evaluated in the validation phase using logistic regression models: estimated coefficients, standard errors, and *P* values of parity scores and other clinical covariates in the validation phase

	Classification tree β(SE) *P*	Tree AdaBoost β (SE) *P*	Support vector machine β (SE) *P*	Random forest β (SE) *P*	Neareast shrunken centroids β (SE) *P*
Parity score[a]	1.799 (0.98) **0.0657**	4.726 (1.77) **0.0075**	3.611 (1.77) **0.0414**	4.122 (1.72) **0.0164**	4.289 (1.30) **0.0010**
BMI	−0.094 (0.13) 0.4640	−0.169 (0.14) 0.2391	−0.155 (0.14) 0.2705	−0.176 (0.14) 0.2242	−0.209 (0.14) 0.1358
Breast density	0.931 (0.72) 0.1931	1.363 (0.81) **0.0923**	1.182 (0.76) 0.1201	1.307 (0.79) **0.0959**	1.071 (0.80) 0.1816
HRT duration	0.037 (0.07) 0.5914	0.037 (0.07) 0.6063	0.022 (0.07) 0.7496	0.022 (0.07) 0.7591	0.048 (0.08) 0.5246
OC use	1.003 (0.67) 0.1342	0.791 (0.70) 0.2602	0.797 (0.67) 0.2308	0.753 (0.68) 0.2662	0.769 (0.78) 0.3247
Smoking duration	−0.013 (0.02) 0.5331	−0.019 (0.02) 0.3707	−0.013 (0.02) 0.5307	−0.015 (0.02) 0.4637	−0.009 (0.02) 0.7087

[a]Parity score is a woman's probability of being parous estimated using the genomic predictors of parity derived in the discovery phase. The significance of the parity scores was evaluated in the validation phase using logistic regression models with FTP as the response variable and parity score, BMI, breast density, HRT duration, OC use, and smoking history as predictors. Statistically significant *P* values are shown in bold.

CLDN10, *CD36*, *PDZD2*, *CLSTN2*, *LAMA4*, *PCDH9*, and *SORBS1* play role in cell adhesion. It is of interest that upregulated genes were more prevalent and consistent than downregulated genes in parous compared with nulliparous women. This suggests that parity results mainly in over-expression of genes involved in breast cell differentiation, organization, and tumor suppression as opposed to down-regulation of genes that might drive development of cancer. The details and biological significance of the genes and pathways differentially expressed in parous versus nulliparous breast will be discussed in a separate paper.

Recently, Asztalos and colleagues (7) examined gene expression in the normal premenopausal human breast, comparing nulliparous, recently parous (0–2 years since pregnancy), and distantly pregnant (5–10 years after pregnancy) age-matched premenopausal women. They analyzed a customized 64-gene set focusing on the genes involved in inflammation, extracellular matrix remodeling, angiogenesis, and estrogen signaling. They reported that 14 of the 64 selected genes were differentially expressed in parous versus nulliparous breast tissues. Compared with nulliparous breast, parous breast had significant upregulation of genes related to inflammation (*CCL21*, *LBP*, *SAA1/2*, and *IGKC*) and down-regulation of genes involved in angiogenesis (*VEGFA*) and estrogen signaling (*ERα*, *PGR*, and *ERBB2*; ref. 7).

There was no overlap between the differentially-expressed genes reported here and those reported by Asztalos and colleagues (7), an inconsistency possibly related to differences in menopausal status, study populations, and laboratory methods. Compared with the hypothesis-driven report by Asztalos and colleagues (7), our study took a comprehensive approach and addressed a much broader list of genes. It also focused on older, postmenopausal women because the objective was to detect long-term gene expression changes, that is, gene expression differences between parous and nulliparous that could be observed many years after the first FTP.

The study's strengths include the formal 2-phase approach for analyses of genomic differences between parous and nulliparous breast, with independent discovery and validation phases but identical procedures throughout the study. The comprehensive gene assessment with micro-array assays is strength of our study. The study focused on healthy subjects attending a mammography clinic and, therefore, representative of the general population of Sweden, a country where mammography is widely accepted. In addition to using rigorous research procedures, stringent criteria were used to control for multiple comparisons and FDR. The laboratory personnel were blinded to samples parity status. In addition, all data analysts were blinded to the parity status of subjects in the validation phase until the parity status predictions were made and the parity scores were derived for the subjects in the validation phase. The consistency of results across 3 different statistical methods used (SAM, SVA, and LRA) strengthened our confidence in the study results.

There were some limitations as well. The study population was restricted to residents of the northernmost part of Sweden, and all participants were of Swedish or Finnish ethnicity. It was felt that these characteristics would be advantageous to avoid gene expression variations resulting from differences in ethnicity rather than parity. The study results, however, should be confirmed in other populations. In addition, although the vast majority of eligible women accepted rather enthusiastically to participate in the study, some women were excluded from the study based on eligibility criteria (34%) or because of RNA degradation, absence of epithelial structures, or an insufficient amount of RNA (32%). Because pregnancy may affect mammographic density, exclusion of women with low-density mammograms may have resulted in differential selection of women at higher, or lower, risk of breast cancer between parous and nulliparous subjects. This is an issue, though, not easily addressable because examination of gene expression cannot be done unless both epithelial structures and sufficient RNA of good quality are present.

We used a FDR of 10%. However, the true FDR corresponding to our list of significant genes is likely to be much lower than 10% because only probes that passed the FDR of 10% in both the discovery and the validation phases were included in the list and those that passed the FDR of 10% in only 1 of the 2 phases were excluded. Additionally, because we were studying normal, rather than pathologic tissues, we *a priori* expected modest effect sizes (e.g., fold change of 1.25) for genomic changes associated with pregnancy in healthy postmenopausal women.

In summary, the results provide initial support to the concept that an FTP induces permanent genomic changes in the breast, thus reflecting the well-known permanent phenotypical changes that follow an FTP. Once further confirmed in additional populations with wider ranges of age, ethnicity, and other characteristics, a well-characterized genomic signature of pregnancy could be used as an intermediate marker for instance to assess potential chemopreventive interventions with hormones mimicking the effects of pregnancy.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Grant Support

References

1. Russo J, Mailo D, Hu YF, Balogh G, Sheriff F, Russo IH. Breast differentiation and its implication in cancer prevention. Clin Cancer Res 2005;11:931s–6s.
2. Russo J, Russo IH. Breast development, hormones and cancer. Adv Exp Med Biol 2008;630:52–6.
3. Russo J, Balogh GA, Heulings R, Mailo DA, Moral R, Russo PA, et al. Molecular basis of pregnancy-induced breast cancer protection. Eur J Cancer Prev 2006;15:306–42.
4. Russo J, Balogh G, Mailo D, Russo PA, Heulings R, Russo IH. The genomic signature of breast cancer prevention. Recent Results Cancer Res 2007;174:131–50.
5. Balogh GA, Heulings R, Mailo DA, Russo PA, Sheriff F, Russo IH, et al. Genomic signature induced by pregnancy in the human breast. Int J Oncol 2006;28:399–410.
6. Balogh GA, Russo J, Mailo DA, Heulings R, Russo PA, Morrison P, et al. The breast of parous women without cancer has a different genomic profile compared to those with cancer. Int J Oncol 2007;31:1165–75.
7. Asztalos S, Gann PH, Hayes MK, Nonn L, Beam CA, Dai Y, et al. Gene expression patterns in the human breast after pregnancy. Cancer Prev Res 2010;3:301–11.
8. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.
9. Anderson K, Hess KR, Kapoor M, Tirrell S, Courtemanche J, Wang B, et al. Reproducibility of gene expression signature-based predictions in replicate experiments. Clin Cancer Res 2006;12:1721–7.
10. Russo J, Reina D, Frederick J, Russo IH. Expression of phenotypical changes by human breast epithelial cells treated with carcinogens in vitro. Cancer Res 1988;48:2837–57.
11. R Development. Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available from: http://www.R-project.org.
12. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
13. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.
14. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
15. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27.
17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
18. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 2007;3: 1724–35.
19. Leek JT, Storey JD. A general framework for multiple testing dependence. Proc Natl Acad Sci U S A 2008;105:18718–23.
20. Qiu X, Xiao Y, Gordon A, Yakovlev A. Assessing stability of gene selection in microarray data analysis. BMC Bioinformatics 2006;7:50.
21. Klebanov L, Yakovlev A. Treating expression levels of different genes as a sample in microarray data analysis: is it worth a risk? Stat Appl Genet Mol Biol 2006;5:Article 9.
22. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 2003;31:2013–35.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Royal Stat Soc B 1995;57:289–300.
24. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72.
25. Vapnik V. The nature of statistical learning theory. New York: Springer; 1996.
26. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New York: Wadsworth; 1984.
27. Freund Y, Schapire R. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference. San Francisco: Morgan Kauffman, 1996:148–56.
28. Breiman L. Random forests. Mach Learn 001;45:5–32.
29. Nikoloski G, Langemeijer SM, Kuiper RP, Knops R, Massop M, Tonnissen ER, et al. Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat Genet 2010;42:665–7.