IJC
International Journal of Cancer

Peripheral blood cells inform on the presence of breast cancer: A population-based case–control study

Vanessa Dumeaux[1,2], Josie Ursini-Siegel[2,3], Arnar Flatberg[4], Hans E. Fjosne[5,6], Jan-Ole Frantzen[7], Marit Muri Holmen[8], Enno Rodegerdts[9], Ellen Schlichting[10] and Eiliv Lund[1]

[1] Institute of Community Medicine, University of Tromsø, Tromsø, Norway
[2] Department of Oncology, Faculty of Medicine, McGill University, Montreal, Canada
[3] Lady Davis Institute for Medical Research, Montreal, Quebec, Canada
[4] NTNU Genomics Core Facility, The Norwegian University of Science and Technology, Trondheim, Norway
[5] Department of Surgery, St. Olavs University Hospital, Trondheim, Norway
[6] Department of Cancer Research and Molecular Medicine, Faculty of Medicine, The Norwegian University of Technology and Science, Trondheim, Norway
[7] Breast Imaging Center, University of North-Norway, Tromso, Norway
[8] Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway
[9] Department of Radiology, Nordland Central Hospital, Bodo, Norway
[10] Department of Cancer, Oslo University Hospital, Oslo, Norway

Tumor–host interactions extend beyond the local microenvironment and cancer development largely depends on the ability of malignant cells to hijack and exploit the normal physiological processes of the host. Here, we established that many genes within peripheral blood cells show differential expression when an untreated breast cancer (BC) is present, and harnessed this fact to construct a 50-gene signature that distinguish BC patients from population-based controls. Our results were derived from a series of large datasets within our unique population-based Norwegian Women and Cancer cohort that allowed us to investigate the influence of medications and tumor characteristics on our blood-based test, and were further tested in two external datasets. Our 50-gene signature contained cytostatic signals including the specific suppression of the immune response and medications influencing transcription involved in those processes were identified as confounders. Through analysis of the biological processes differentially expressed in blood, we were able to provide a rationale as to why the systemic response of the host may be a reliable marker of BC, characterized by the underexpression of both immune-specific pathways and "universal" cell programs driven by MYC (*i.e.*, metabolism, growth and cell cycle). In conclusion, gene expression of peripheral blood cells is markedly perturbed by the specific presence of carcinoma in the breast and these changes simultaneously engage a number of systemic cytostatic signals emerging connections with immune escape of BC.

Epidemiology

Cancers are not simply autonomous masses of cells; they secrete soluble factors that elicit systemic responses from the host, and the host responses, in turn, affect cancer cells.[1–4] Several studies support the notion that tumors act systemically to modulate overall cancer progression[5–7] and that cancer development largely depends on the ability of malignant cells to hijack and exploit the normal physiological processes of the host.[3,4]

Key words: breast cancer, population-based case–control study, tumor–host interactions, blood gene expression profiling

Abbreviations: APP: antigen processing and presentation; BC: breast cancer; CC: case–control series; DAVID: Database for Annotation, Visualization and Integrated Discovery; FDR: false discovery rate; GSVA: gene set variation analysis; NBC: naive Bayes classifier; NOWAC: Norwegian Women and Cancer study; NK: natural killer; PBMC: peripheral blood mononuclear cells

Additional Supporting Information may be found in the online version of this article.

VD and EL have filed a patent from the University of Tromso and Norinnova for the identified blood-based gene-list assay diagnostic of breast cancer

Some of the data in this article are from the Cancer Registry of Norway that is not responsible for the analysis or interpretation of the data presented

Grant sponsor: European Research Council; **Grant number:** ERC-2008-AdG 232997
DOI: 10.1002/ijc.29030
History: Received 20 Dec 2013; Accepted 2 June 2014; Online 14 June 2014

Correspondence to: Vanessa Dumeaux, McGill University, Bellini Building, Room 434, 3649 Sir William Osler, Montreal, Canada QC H3G 0B1, Tel.: ***514-398-5928, Fax: 514-398-3387, E-mail: vanessa.dumeaux@mcgill.ca

What's new?

Blood cells are dynamic warehouses of information. In the case of cancer, studies have indicated that blood cells house genetic signatures related to solid tumors. In the present study, genes in peripheral blood cells were found to be differentially expressed in women with untreated breast cancer, enabling the development of a 50-gene signature capable of identifying women with the disease. The gene signature included signals specific to immunosuppression. The association of breast cancer with the underexpression of immune-specific pathways and with MYC-driven "universal" cell programs may explain the systemic response of the host.

Despite the emerging appreciation for cancer–host interaction in cancer, our understanding of how the host responds to cancer signals and in turn affects tumor progression and prognosis is rudimentary. There is growing evidence that gene expression profiling of peripheral blood cells is a valuable tool for assessing gene signatures related to solid tumors.[8–11] Several groups including ours have defined intra- and interindividual variability of blood gene expression in healthy individuals[12,13] and established standardized procedures for blood sample collection and gene expression profiling.[14,15] In this study, we selected a large number of breast cancer (BC) patients and women representative of the general population matched on birth year and time of follow-up in the Norwegian Women and Cancer study (NOWAC).[16,17] To the best of our knowledge, this is the first study in this domain that accurately represents the actual situation of identifying BC patients from women representative of the general population. We followed a strict experimental design where each case sample was processed with an age-matched control throughout all steps of the laboratory procedures from RNA amplification to hybridization. Paired analyses within two training sets and one validation set were carried out to ablate any technical bias and help ensure generalizability of the results. The accuracy of the identified 50-gene blood signature was first evaluated with respect to lifestyle exposures and tumor characteristics. The signature was then further tested in two additional external gene expression datasets using profiles of blood cells or isolated immune cells from BC patients, and women with suspect mammograms or diagnosed with benign breast diseases. The behavior of our multigene signature was also evaluated in other cancer types to assess the specificity of the system to BC. Finally, we investigated the molecular processes involved in the systemic response of the host and provide a rationale as to why blood-based gene expression may harbor a reliable signal of the presence of BC.

Methods

The NOWAC study

The NOWAC study consists of 172,471 women 30 to 70 years of age at recruitment from 1991 to 2006 who answered one to three questionnaires on diet, medication use and lifestyle.[18] Ten of the largest Norwegian hospitals participate in collecting blood and tumor tissue from incident BC cases.[16] In collaboration with the Norwegian Breast Cancer Group, every woman born between 1943 and 1957 participating in the NOWAC study who is admitted to a collaborating hospital for a diagnostic biopsy or for surgery of BC was asked to donate, before surgery and treatment, a tumor biopsy and two blood samples, one collected into PAXgene^{TM} tube (PreAnalytiX GmbH, Hembrechtikon, Switzerland) for gene expression analysis and another in a citrate tube. Participants were also asked to answer a two-page questionnaire eliciting information mainly on current use of hormones and medications, alcohol and smoking habits. Biological samples were then mailed overnight for biobanking at −70°C in Tromsø. In parallel, five controls were approached for each BC case in order to obtain blood samples from at least two controls per case. The controls were drawn at random but matched by time of inclusion in the NOWAC cohort and birth year. The human biological material has been approved by Regional Committees for Medical and Health Research Ethics in Norway and is in accordance with the Norwegian law on biobanking.

In 2009, 96 blood samples from cases and two matched controls for each case were selected from the postgenome biobank (CC1). In 2010, 63 blood samples received within 4 days after blood collection from cases and one matched control for each case were selected from the postgenome biobank (CC2). In 2011, 90 blood samples received within 4 days after blood collection from cases and one matched control for each case were selected from the postgenome biobank (CC3).

Microarray data acquisition

To control for technical variability such as different lot variations of reagents and kits, day to day variations, microarray production batches and effects related to different laboratory operators, each case was grouped with one corresponding matched control through RNA extraction (except in CC1 where RNA extraction was run randomly), amplification and hybridization. Total RNA from cases and matched controls for each case were isolated using the PAXgene Blood miRNA Isolation Kit according to the manufacturer's manual at the NTNU Genomics Core Facility in Trondheim, Norway. RNA quantity and purity was assessed using the NanoDrop ND-8000 spectrophotometer (ThermoFisher Scientific, Wilmington, Delaware) and Agilent bioanalyzer (Palo Alto, CA), respectively. RNA amplification was performed in 96 plates using 300 ng of total RNA and the Illumina® TotalPrep^{TM}-96 RNA Amplification Kit (Ambion, Austin, TX). Cases and controls included in CC1 and CC2 were run on the IlluminaHumanAWG-6 version 3 expression bead chips.

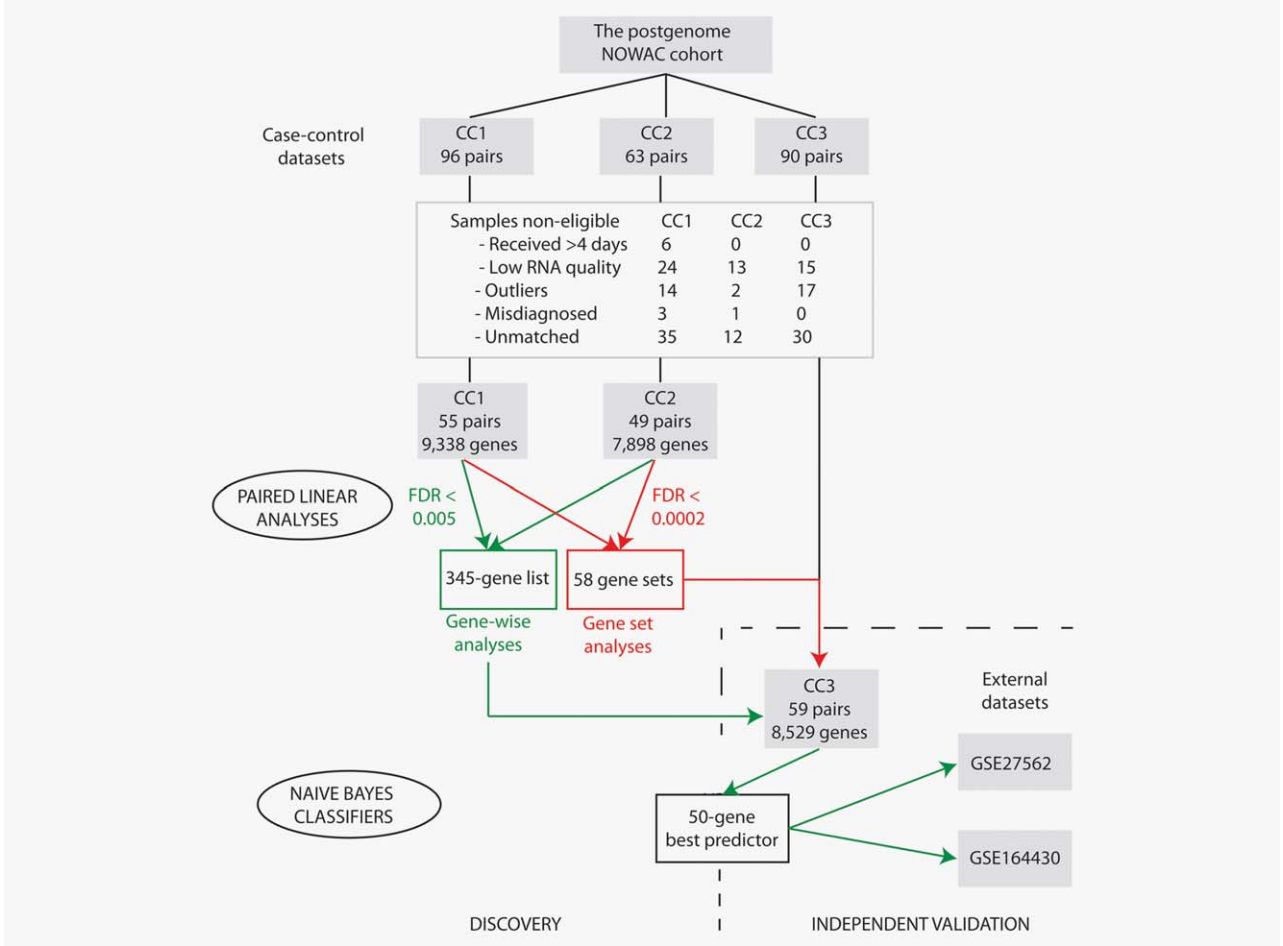

Figure 1. Study flow chart. Sample exclusions are shown for the three independent case–control datasets from the Norwegian Woman and Cancer (NOWAC) study (CC1, CC2 and CC3). Paired linear analyses were conducted to identify single genes (False Discovery Rate, FDR < 0.005) and gene sets (FDR < 0.002) differentially expressed across BC case–control pairs. Prediction of the presence of BC in CC3 based on the expression of the 345 genes differentially expressed in both CC1 and CC2 was conducted using a naive Bayes classifier. Fifty genes were further selected among the 345-gene list and validated in two external datasets from NCBI's Gene Expression Omnibus (including gene expression profiles of peripheral blood mononuclear cells (PBMC) from BC patients, patients with benign breast diseases, controls, gastrointestinal and brain cancer patients[28] (GSE27562) and gene expression profiles of peripheral blood cells from BC patients and controls with suspect mammograms[9] (GSE164430). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Cases and controls included in CC3 were run on the IlluminaHumanHT-12 version 4 expression bead chips. GenomeStudio from Illumina (San Diego, CA) was used to assess the quality of each array.

Microarray data preprocessing

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

Data preprocessing was done identically for all three independent datasets (Fig. 1). We excluded samples received for more than 4 days after collection and samples with low RNA quality (RIN < 7). The datasets were trimmed of samples found misdiagnosed after update from the cancer registry or found outliers as called by the lumiR package.[19] More precisely, outliers were identified when their euclidean distance to the cluster center was larger than twice the median distances to the center. The cluster center was defined by the average of all samples after removing 10% samples farthest away from the center. Finally, resulting unmatched samples from the above exclusion procedures were excluded from the analyses. Preprocessing of the microarray data was performed in each dataset separately using the lumiR package. One probe was defined as present if its intensity was significantly different from the background intensity (p value < 0.05), resulting in the analysis of 48,803 probes in CC1 and CC2, and 47,323 probes in CC3. We excluded all probes that did not have an expression value in at least 70% of the samples leading to 13,460, 10,341 and 12,519 probes in CC1, CC2 and CC3, respectively. Variance stabilization[20] was performed and the data were normalized by quantile normalization. We used the reannotation pipeling for Illumina arrays[21] version 3 for CC1 and CC2, and version 4 for CC3. Intensities of probes

with similar gene symbols were averaged leading to 9,338, 7,898 and 8,529 unique gene symbols in CC1, CC2 and CC3, respectively. Microarray data have been deposited at the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) accession number EGAS00000000134.

Technical variability

Amplification date was associated with our gene expression profiles but this latter remains independent of BC status since each case was amplified with its control in the same round and paired analyses were conducted. In CC1, RNA extraction of blood samples was done by random in CC1 and date of RNA extraction was significantly associated with disease status (χ^2 test: p value = 0.02). Supporting Information Additional file 1 shows the ordering of the samples based on the expressions of the top quintile of most variable genes in CC1. Overall, the variable coding for six different dates of RNA extraction did not seem to be strongly associated with blood gene expression profiles and therefore its effect has not been adjusted for in further analyses.

Gene-wise paired linear analysis

To identify single gene differentially expressed between cases and controls, we conducted paired gene-wise linear analysis with application of empirical Bayes method[22] implemented in the software package Limma in each dataset. False discovery rate (FDR)[23] was calculated to adjust for multiple testing.

Prediction of the presence of BC

Despite the fact that the independence assumptions are inaccurate, the naive Bayes algorithm[24] alleviates problems stemming from the curse of dimensionality and was implemented as the class prediction method. Indeed, naive Bayes is a well-established learning approach used for gene expression studies due to its simplicity, interpretability, and has been shown to have performance similar to, and sometimes exceeding, those of much more complex and approaches.[25] Gene selection was performed among the genes commonly differentially expressed in CC1 and CC2 by the presence of BC (paired linear analysis FDR < 0.005; N = 345, Fig. 1) in order to optimize robustness of our signature across datasets. One naive Bayes classifier (NBC) was first built using all genes expressed in CC3 (N = 341) and tested by leave-one-out cross validation (LOOCV). Prediction accuracy is the fraction of true *versus* all predictions derived from the posterior probability generated by the NBC. The significance of a NBC was tested by a Fisher's exact test measuring the strength of association between observed and predicted disease status. Our classifier significance was then compared with the background distribution of significance obtained from 100,000 NBCs including 341 genes randomly chosen from the 8,529 genes expressed in CC3. Similarly, we constructed NBCs including several subset sizes (10, 25 or 50 genes) of the 341 overlapping genes expressed in CC3 where a predictor of size 50 appeared the most appropriate (Supporting Information Additional file 4A). We also tested 100,000 NBCs of size 50 randomly chosen from larger lists of genes commonly differentially expressed in CC1 and CC2 by the presence of BC (N = 565 genes with FDR < 0.01 and N = 1,426 with FDR < 0.05) but did not witness any improvement in terms of the predictors significance (Supporting Information Additional file 4B). The "best" 50-gene predictor was empirically selected among the 100,000 predictors built using 50 genes among the 341 expressed in CC3 based on its statistical significance in predicting the presence of BC (Fisher's test). Its significance was further compared to the background distribution of 100,000 random 50-gene NBCs.

In all three datasets, we investigated using the student or chi-square tests whether RNA quality quantified by the RIN value, individual or exposure variables such as age, BMI or smoking status, and the use of menopausal hormone therapies or other specific medications could explain the misclassification of controls (*i.e.*, false positives) and cases (*i.e.*, false negatives). In the same manner, we investigated whether tumor receptor status (estrogen and progesterone receptor) or stage was associated with misclassified cases by our 50-gene predictor in all three datasets. None of these variables were associated with misclassification of controls or cases except for the use of specific medications by controls in CC3 (see Results). Of note, BC were mostly ER positive and of stage I or II (Supporting Information Additional file 1). Finally, we investigated perturbagen signatures ($n > 10$) from the connectivity map[26] significantly enriched in our 50-gene predictor.

Independent validation of the "best" 50-gene predictor was conducted using two additional external datasets deposited in NCBI's Gene Expression Omnibus[27] including gene expression profiles of peripheral blood mononuclear cells (PBMCs) from BC patients, patients from benign breast diseases, controls, gastrointestinal and brain cancer patients[28] (GSE27562) and gene expression profiles of whole blood cells from BC patients and controls with suspect mammograms[9] (GSE164430; Fig. 1). We represented genes by symbols assigned by the HUGO Gene Nomenclature Committee.

Functional clustering and pathway analysis

Functional clustering of the gene lists associated to BC diagnosis was performed with the Database for Annotation, Visualization, and Integrated Discovery (DAVID)[29] at http://david.abcc.ncifcrf.gov/. For each functional cluster, we selected the terms with FDR < 0.1 and calculated the median fold enrichment and FDR.

Gene set analysis

Enrichment scores for pathways (size; min = 5 and max = 500) included in release 3.0 of the C2 (curated gene sets) and C5 (Gene ontology gene sets) subcollections of the Molecular Signatures Database[30] were calculated for each sample using the GSVA R package. We build gene sets specific of immune cell subtypes using CD markers of no more

than three immune cell subtypes[31] and transcripts specifically overexpressed in each differentiated immune cell subtypes,[32] and calculated in the same manner enrichment scores for each sample. We tested whether there is a difference between the enrichment scores across case–control pairs using paired linear analysis as described previously.[22]

The significant contribution of each gene or sample within a pathway of interest to the overall test statistics for differential expression was estimated by the global test covariate analysis.[33] Covariate and subject plots in the globaltest R package estimates the contribution of each (cluster of) gene(s) (covariate) or sample (subject) to the overall test statistics for differential expression plotting the p values of the tests of individual component of the alternative. Samples were ordered by decreasing order. Genes are ordered in a hierarchical clustering using correlation as a distance measure. The hierarchical clustering graph induces a collection of subsets of the tested covariates between the full set that is the top of the clustering graph and the single covariates that are the leaves. Inheritance procedure on all $2k-1$ sets controls the family-wise error rate while taking the structure of the graph into account.[34]

Results

Blood-wide transcriptional signal of BC and its potential to detect BC

Disease status was associated with substantial differences in blood gene expression profiles across the case–control pairs included in CC1 ($p = 6 \times 10^{-8}$, global test). This blood-wide signal of BC is illustrated by the grouping of samples according to disease status based on the expression of the most variable genes (Supporting Information Additional file 2). We identified 3,479 genes exhibiting significant differences in expression within the BC case and control pairs (FDR < 0.005; paired linear analysis) with a relatively low median absolute value of fold-change equal to 1.13. This indicates that gene expression changes associated with BC are of relatively low amplitude but consistent and ubiquitous in peripheral blood cells.

To test whether these findings were replicable in an independent data set (CC2), we investigated blood gene expression profiles from an additional 49 pairs of BC cases and controls (Fig. 1). In total, 418 of the 7,898 genes passing quality controls were differentially expressed in CC2 with a FDR < 0.005, of which 345 were also differentially expressed in CC1 ($p = 3 \times 10^{-60}$, hypergeometric test; Fig. 2*a*, Supporting Information Additional file 3). Remarkably, the directionality of differential expression between BC cases and controls of all 345 overlapping genes was conserved between datasets (Fig. 2*b*). When patients were ranked according to the sum of expression over the 345 overlapping genes, the majority of blood samples from BC cases were segregated from controls in both datasets (Fig. 2*c*).

Using both CC1 and CC2 to select genes differentially expressed in blood cells from BC patients compared to controls ($N = 345$, Supporting Information Additional file 3), we built a predictor using the 341 genes expressed in CC3 (four genes were not present in CC3) and accurately predicted disease status in this validation dataset ($p = 8.7 \times 10^{-5}$; Fisher's test; Fig. 2*d*). Of note, amplified RNA from the blood samples in CC3 was hybridized using a different version of the Illumina array system. The "best" 50-gene predictor (Fig. 2*d*, Supporting Information Additional file 3) chosen among the 341 significant genes had an accuracy of 72.9% to predict the presence of BC in the validation dataset (sensitivity = 83.1% and specificity = 62.7%; $p = 3.0 \times 10^{-9}$; Fisher's test). Notably, gene expression signatures from the connectivity map[26] associated with histone deacetylase, Hsp90, tyrosine kinase and immune response inhibitors were positively enriched with our 50-gene predictor (Supporting Information Additional file 5). A significant proportion of controls misclassified as cases in CC3 (36.4%) were currently using either a selective serotonin reuptake inhibitor (ATC N06AB) or a selective beta-blocking agent (ATC C07AB). Both drugs associated with misclassified controls in CC3 were previously found to inhibit the expression of T-cell and adaptive immunity-related genes.[35,36] This may explain the lower specificity of our 50-gene predictor in CC3. Overall, this indicates that our 50-gene predictor contains cytostatic signals including the specific suppression of immunity and that medications influencing transcription involved in those processes can be confounder of the blood-based signal associated with the presence of BC.

To further validate the results, we investigated whether we were able to predict BC diagnosis in two external datasets deposited in NCBI's Gene Expression Omnibus[27] including gene expression profiles of PBMCs from BC patients, patients with benign breast diseases, controls, gastrointestinal and brain cancer patients[28] (GSE27562), and gene expression profiles of peripheral blood cells from BC patients and controls with suspect mammograms[9] (GSE164430; Fig. 1). In the PBMC dataset, our 50-gene predictor was able to accurately predict the presence of BC compared to controls (91.5% accuracy, Supporting Information Additional file 6). This indicates that our diagnostic profile for BC identified from peripheral blood cells is found in isolated immune cells including monocytes, T-cells, B-cells and natural killer (NK) cells. All PBMC samples from other cancer types were not predicted as BC, which indicates that our predictor is specific for carcinoma in the breast. Since our predictor was not trained to differentiate malignant BC from benign breast diseases, we obtained significantly lower accuracy when we included those samples (63.4% accuracy; Supporting Information Additional file 6). The expression of only 33 genes of our 50-gene predictor were available in the second dataset (GSE164430) although we were able to significantly predict BC diagnosis compared to women with suspect screening mammograms ($p = 0.008$; Fisher's test, Supporting Information Additional file 7). In conclusion, our blood-based gene expression analysis produced uniquely robust and reproducible results across microarray platforms and external datasets

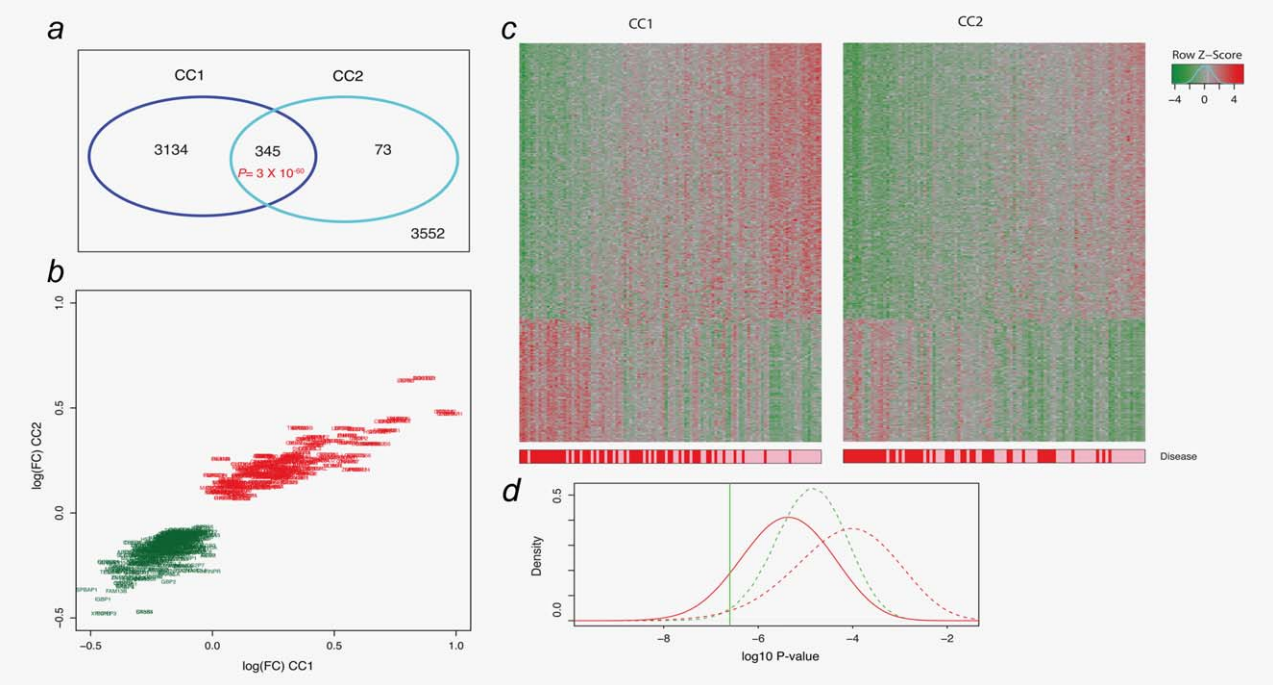

Figure 2. Gene expression changes in peripheral blood cells of breast cancer (BC) patients compared to controls. (*a*) Venn diagram depicting the overlap between genes differentially expressed in peripheral blood cells of BC patients compared to controls in the primary (CC1) and the secondary (CC2) dataset. Differential expression was assessed at an FDR < 0.005 by the paired linear analyses in CC1 and CC2. The significance of overlap between the two gene lists was calculated using the hypergeometric test. (*b*) Expression fold changes for the 345 overlapping genes differentially expressed in CC1 and CC2. Log fold-changes (log FC) in CC1 are plotted on the x-axis against the log FCs for the same genes in CC2 on the *y*-axis. Genes in green are underexpressed by the presence of BC in both data sets. Genes in red are overexpressed by the presence of BC in both data sets. (*c*) Ordering of blood samples from BC cases (in red) and controls (in pink) according to the sum of expression over those 345 overlapping genes. Heat map colors represent mean-centered fold change expression in log-space. (*d*) Significance of naive Bayes classifiers in the validation dataset (CC3) calculated using Fisher's test. Vertical green line represents the significance of a naive Bayes predictor based on the expression of the 345 overlapping genes. The dotted green line represents the distribution of significances that can be obtained from 100,000 naive Bayes predictors built using 345 random genes present in CC3 ($N = 8{,}529$). Plain red line represents the distribution of significances that can be obtained from 100,000 predictors built using 50 genes among the 341 expressed in CC3 of the 345 overlapping genes. Dotted red line represents the distribution of significances that can be obtained from 100,000 predictors built using 50 random genes present in the dataset ($N = 8{,}529$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to specifically detect BC from population-based controls, warranting further analysis of the processes found deregulated by the presence of a BC in peripheral blood cells including PBMC.

Pathway and gene set analyses

Functional clustering showed that our 345-gene list was enriched for gene ontology categories related to apoptosis, RNA binding, spliceosome/RNA splicing, protein synthesis, RNA metabolism, transcriptional regulation, cell cycle, metabolism and signal transduction (Table 1). Expert curated functional annotation revealed additional grouping of genes involved in immune processes, cell growth/proliferation, cytoskeletal regulation and protein and cell metabolism (Supporting Information Additional file 3).

To further investigate how BC affects gene expression in blood, we performed gene set variation analysis (GSVA) in CC1 and CC2 datasets and validate the results in CC3 (Fig. 1). We found 58 gene sets overlapping between the top 200 gene sets differentially expressed across case–control pairs in CC1 and CC2 (FDR $< 2 \times 10^{-4}$). Although we previously identified a confounding factor with the current use of specific drugs by controls in CC3, 45 of the 58 significant gene sets overlapping in CC1 and CC2 were validated in CC3 (FDR < 0.15), and showed remarkably comparable enrichment scores according to disease status in all three datasets (Fig. 3). GSVA revealed similar processes seen after functional clustering of our 345-gene list including transcription and cytoskeletal regulation, cell cycle, apoptosis and metabolism pathways, but also identified additional gene signatures notably involved in antigen processing and presentation (APP) and MYC target genes (Fig. 3, Supporting Information Additional file 8).

BC and the host's immune system

Reduced expression of APP pathway in blood cells of BC patients was the most direct evidence that the presence of BC affects peripheral immune effector cells (Fig. 3, Supporting Information Additional file 8). We investigated the overlapping core genes (multiplicity corrected p value <0.1) driving the observed association of the APP pathway with the

Table 1. Functional annotation clustering of significant enrichments associated with the 345-gene list (false discovery rate, FDR < 0.10) differentially expressed across breast cancer case–control pairs

Annotation terms, *N*	Annotation cluster (keywords)	Genes,[1] *N*	Fold enrichment[2]	FDR[2] (%)
8	Cell death, apoptosis	35	2.7	0.01
11	Regulation of apoptosis	37	2.3	0.01
5	RNA catabolic process	9	10.4	0.07
15	RNA binding, protein synthesis, translation, ribosome	54	3.9	0.05
2	RNA binding protein, RRM	35	3.7	3.80
11	RNA processing, splicing, spliceosome	29	3.1	1.73
3	Protein kinase binding, enzyme binding	23	3.2	3.27
6	Protein phosphatase activity, manganese	11	5.0	2.28
2	Response to inorganic substance or metal ion	12	3.2	6.02
2	Ribosome, ribonucleoprotein biogenesis	11	3.6	3.48
7	Regulation of transcription, macromolecule metabolic process, nitrogen compound, RNA pol II	41	3.7	2.53
3	Nucleotide or ATP binding	67	1.6	6.82
1	Acetylated amino end	23	4.6	2.44
1	Endoplasmic reticulum	23	1.8	9.00
2	Cell cycle	32	2.1	0.53
1	Myristylation	5	7.4	6.06
2	Generation of precursor metabolites and energy, glycolysis	16	6.0	5.25
1	Oxidoreductase activity, acting on sulfur group of donors	5	6.7	8.72
1	Ras protein signal transduction	8	3.8	8.42

[1]Number of genes from the 345 gene list involved in the corresponding processes.
[2]Median value across all significant annotation terms included in the cluster.

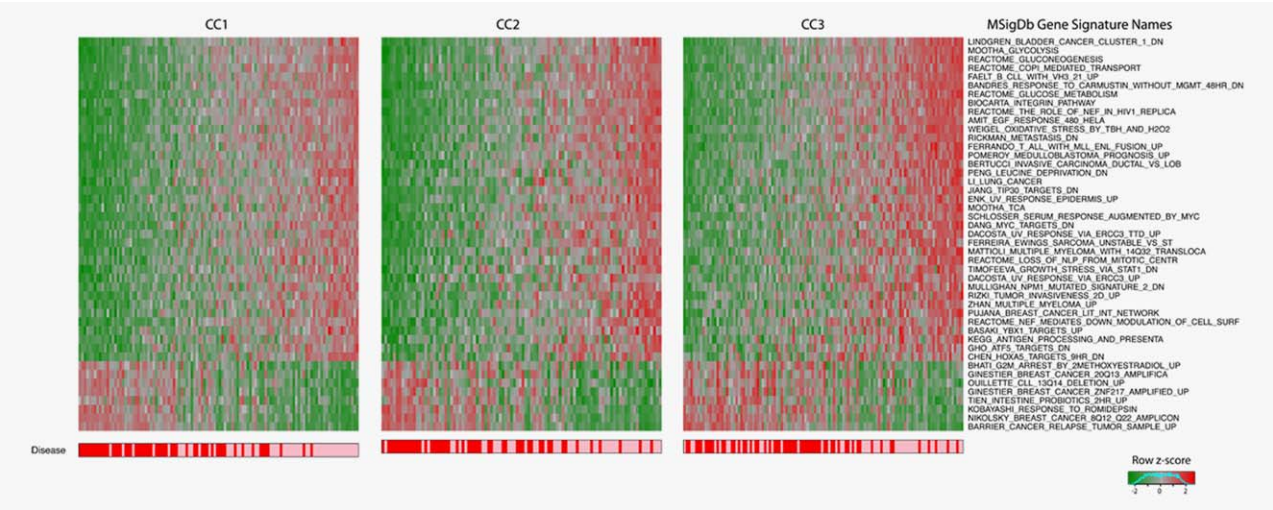

Figure 3. Ordering of blood samples based on the enrichment scores of the 45 significant gene sets differentially expressed between breast cancer cases (in red) and controls (in pink) in the primary (CC1), secondary (CC2) and validation (CC3) case–control series. Heat map colors represent mean-centered fold change enrichment score in log-space. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

presence of BC within CC1 and CC2 datasets (Fig. 4*a*, Supporting Information Additional file 9). All core genes that are part of the MHC class II pathway were underexpressed in blood samples from BC patients compared to controls including the interferon gamma-inducible protein 30 and cathepsin S involved in the endocytic generation of MHC class II-restricted epitopes as well as *CD74* involved in the formation and transport of MHC class II protein, and *CD4* a co-

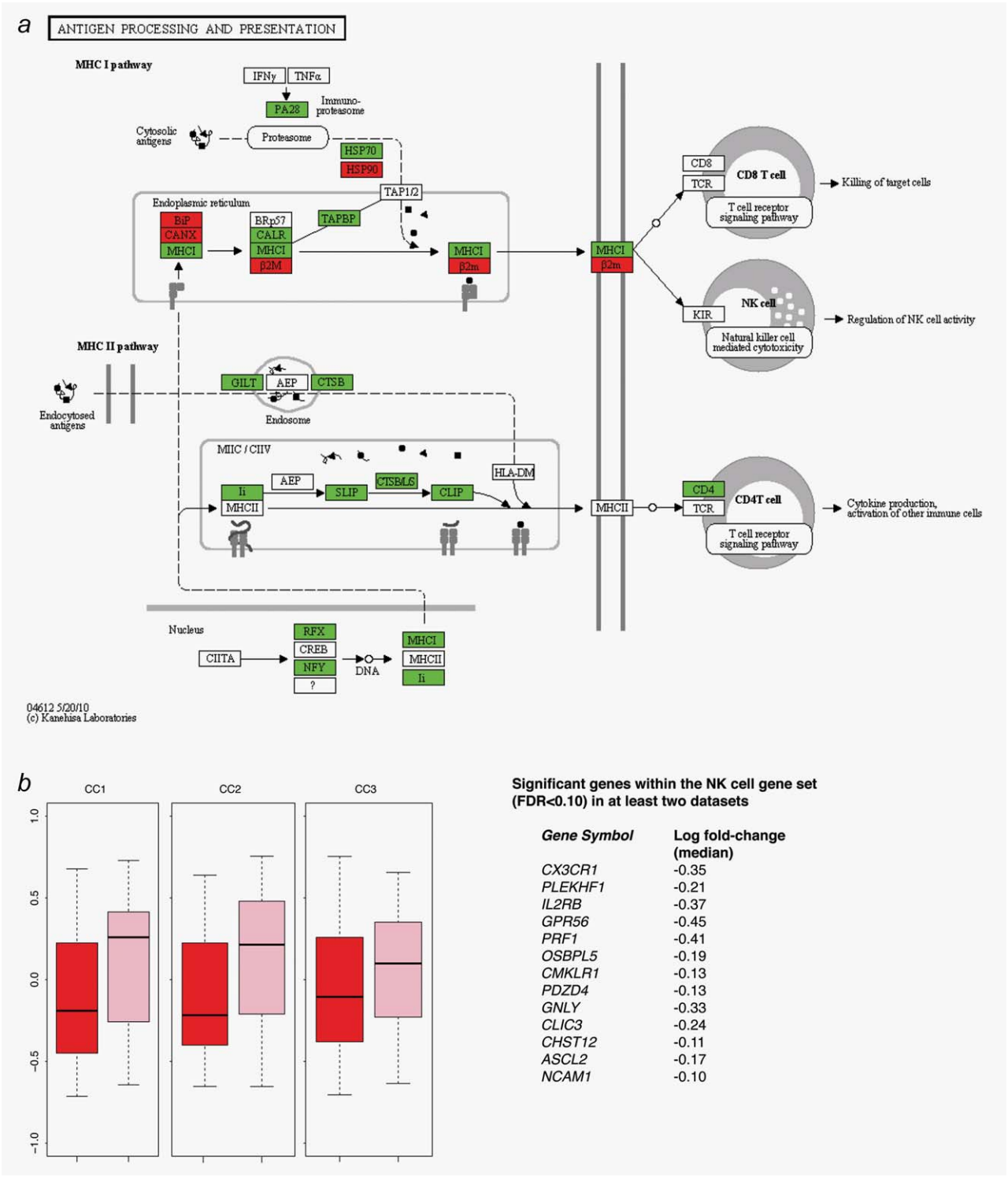

Figure 4. Gene set variation analysis of the antigen processing and presentation pathway (APP) and the natural killer (NK) cell gene set. (*a*) The APP from KEGG. Overlapping core genes driving the observed association of the APP with the presence of BC within CC1 and CC2 are colored according to their over- (red) or under- (green) expression in BC patients. (*b*) Boxplot indicating the enrichment scores from gene set variation analysis for the NK cell gene set associated with the gene expression profiles from BC patients (red) and controls (pink) included in CC1, CC2 and CC3 (left). List of genes included in the NK gene set significantly associated to disease status in paired linear analysis with FDR < 0.10 in at least two of the three datasets and their corresponding median fold-changes over all datasets (right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

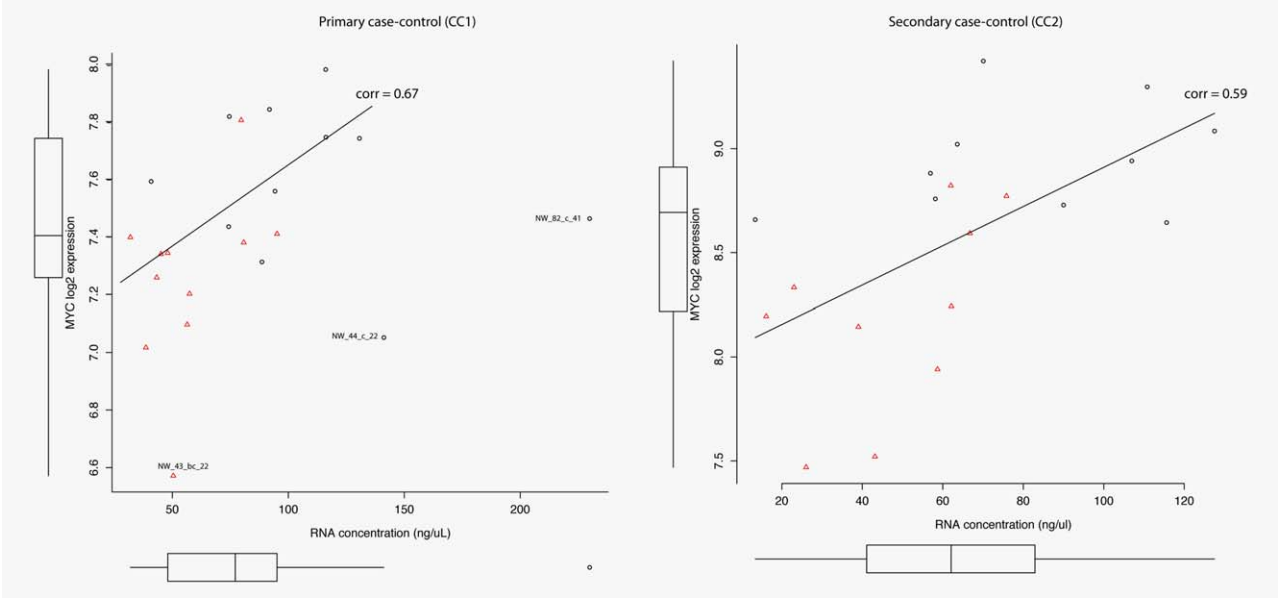

Figure 5. RNA concentrations of the top quintile blood samples contributing the most to the differential enrichment of the MYC gene set between BC patients (in red) and controls (in black) according to the expression of *MYC*. Spearman correlation (corr) is given for each linear regression line in the primary (CC1) and secondary (CC2) case–control series. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

receptor that assists the T-cell receptor (TCR). Within the MHC class I pathway, *PSME3* of the immune-proteasome was defined as a core gene in the APP pathway underexpressed in BC patients. In addition, three genes coding for the proteasome (*PSMB2, PSMB10* and *PSMD1*) within our 345-gene list (Supporting Information Additional file 3) were underexpressed in blood cells of BC patients compared to controls. Also, core genes directly involved in peptide loading onto MHC class I molecules (*TAPBP* and *CALR*) and genes encoding for the heat shock protein 70 upstream of TAP were underexpressed in BC patients.

Finally, genes specific to NK cells were consistently underexpressed within the set and in samples from BC patients compared to controls (Fig. 4*b*, Supporting Information Additional file 10). Consistent with this finding, the NK cell-mediated cytotoxicity pathway from KEGG was significantly underexpressed in blood samples from BC patients compared to controls in CC1 and CC2 (mean FDR = 0.004; paired linear analysis).

Underexpression of Myc and decreased metabolism, growth and proliferation in peripheral blood cells from BC patients

We found a decrease in gene set expression of the targets regulated by Myc according to Myc Target Gene Database[37] in blood cells from BC patients (Fig. 3, Supporting Information Additional file 8). The Myc targets defined as core genes were involved in global gene regulatory networks with specific influence on cell growth and proliferation (*ILK, ARPC4, PP2R4, ERBB2* and *CEBPA*).

Total RNA levels were compared between groups to investigate the effect of Myc recently reframed as a general amplifier of gene expression for all active genes.[38] Samples from BC cases had lower RNA levels than blood samples from matched controls (p value $= 8 \times 10^{-6}$; logistic regression). RNA degradation only, measured by the RIN value, could not explain the decrease in RNA yield in blood samples from BC patients compared to controls (data not shown). In a further attempt to distill the essentials of Myc action in peripheral blood cells of BC patients, the binding and expression changes among another set of genes accepted as bona fide Myc targets were highlighted.[38,39] We compared RNA concentration and the expression of *MYC* in the top quintiles of samples contributing the most to the enrichment's significance of the bona fide Myc target gene set (Supporting Information Additional file 11). The expression of *MYC* was significantly correlated to the RNA concentration with a significant underexpression of *MYC* in blood samples of BC patients compared to controls (Fig. 5). Consistent with this, most differentially expressed transcription factors, genes that are part of the general transcription machinery and involved in chromatin remodeling were underexpressed in blood cells from BC patients compared to controls confirming that steady state transcription rates are reduced (Supporting Information Additional file 3).

Even a modest reduction in Myc may be sufficient to deprive cells of the net anabolic, metabolic and mitogenic impulse necessary to sustain growth and proliferation. In our study, we observed a lowered cell metabolism with an overall underexpression of glycolysis and glucose metabolism pathways in blood cells of BC patients (Table 1, Fig. 3, Supporting Information Additional file 8). In accordance with this, two genes promoting autophagy (*ATG12* and *VPM1*) in our 345-gene list were overexpressed in blood cells of BC

patients. Consistent with a decrease in protein synthesis and consequently cell growth, three components of the eukaryotic initiation factor 4 complex (*EIF4A1*, *EIF4A3* and *EIF4H*) were significantly underexpressed in BC patients compared to controls. Furthermore, several genes involved in ribosomal biogenesis (*GAR1, SURF6* and *RRS1)* were underexpressed while several small *(RPS3A* and *RPS29*) and large (*RPL4, RPL5, RPL7, RPL11, RPL15, RPL21* and *RPL41*) ribosomal proteins (RPs) were overexpressed in blood cells from BC patients compared to controls. Finally, several genes (*TERF2, CKAP5, CUL4B, MCM3, HBP1, NUDC, CTCF, TUBB, USP9X* and *H2AFX*) significantly underexpressed in peripheral blood cells of BC patients were involved in regulating cell cycle (Table 1). GSVA pointed to processes involved in mitosis checkpoint (centrosome maturation, loss of Nlp from mitotic centrosomes and G2-M transition, Fig. 3, Supporting Information Additional file 8). Finally, the integrin signaling pathway involved in cellular shape, mobility and progression through the cell cycle was underexpressed in blood samples from BC patients compared to controls (Fig. 3, Supporting Information Additional file 8).

Discussion

Gene expression changes associated with BC were of relatively low amplitude but consistent and ubiquitous in peripheral blood cells, and clearly identified in isolated immune cells (*i.e.*, PBMC). Our 50-gene signature contained cytostatic signals including the specific suppression of the immune response and medications influencing transcription involved in those processes were confounder of the blood-based signal associated with the presence of BC. Our blood-based gene expression analysis produced uniquely robust and reproducible results across microarray platforms and external datasets to specifically detect BC cases from population-based controls.

Together, our findings uniquely indicate that the presence of BC is associated with systemic immunosuppression by underexpression of several immune-specific pathways (*i.e.*, APP, NK cell-mediated immunity) and several MYC-driven "universal" cell programs (*i.e.*, cell metabolism, growth and proliferation). Mechanisms that regulate APP alter the form and the quantity of the epitopes that are presented by the MHC molecules for immune recognition and can dictate tumor immunogenicity.[40] Our study uniquely shows specific alterations in the APP associated with systemic immunity and confirms some mechanisms previously identified within the tumor and its microenvironment including down-regulation of MHC I molecules, proteasome subunits and transport associated with antigen presentation (TAP and Hsp70) and MHC-peptide complexes.[41] Reduced expression of MHC I molecules may induce NK-cell cytotoxicity although we observed concurrent qualitative impairment of NK-cell mediated immunity in blood samples from BC patients. Remarkably, one epidemiological study has previously associated low peripheral blood NK-cell cytotoxic activity with increased cancer risk.[42] Finally, we observed overall underexpression of genes involved in MHC-II-restricted antigen presentation as well as *CD4* necessary to TCR-mediated activation of helper T cells.

Our study is the first to show that an overall decrease in RNA levels in blood cells of BC patients compared to population-based controls correlates with *MYC* expression in certain BC patients. It is thought that Myc is ubiquitously expressed in proliferating cells[38] where it controls RNA processing, ribosome biogenesis, protein synthesis, metabolism and the cell cycle for normal cell growth and proliferation.[38] Importantly, all those processes were found underexpressed in blood cells from BC patients compared to controls. Plasma levels of enzymes involved in glucose, lipid and amino acid metabolism were previously found altered during tumor development in mice[43] confirming that the presence of a tumor triggers systemic metabolic dysregulation. While cell metabolism was limited in blood cells of BC patients, some genes activating autophagy were found overexpressed to provide a source of ATP. In our study, the rates of protein synthesis *via* translation initiation/elongation and ribosome biogenesis were decreased in blood cells from BC patients where RPs accumulates possibly due to defects in ribosome assembly.[44] A fine regulation of the cell cycle is also required to maintain cell homeostasis, although expression of several genes involved in cell cycle and processes related to mitosis checkpoint were differentially expressed in blood cells of BC patients compared to controls.

Our results suggest that processes found deregulated in blood cells reflect a deficit in immune functions of BC patients. Although we did not isolate effector immune cells from blood, we observed a concerted decrease in expression of genes involved in crucial functions for antitumor immune response (*e.g.*, APP, NK cell-mediated cytotoxicity).[41,45] Furthermore, the observed cytostatic signals in blood cells of BC patients were correlated with gene expression profiles associated with exposure to immunosuppressive medications. Of note, the changes in blood gene expression could represent altered blood cell composition or changes in gene expression from distinct cellular populations. Although the effect of tumor development on peripheral immune cells count has not been clarified, impairment of APP, activation of negative costimulatory signals and production of immunosuppressive factors (or cells) by the tumor may induce lymphopenia in cancer patients which has been found associated with patient prognosis.[46–48] This study first points to systemic molecular dysfunction in the host's immune response to the presence of BC compared to population-based controls that may reflect tumor immune escape.

Some questions that remain unanswered are how *MYC* expression is repressed in peripheral blood cells by the presence of a specific BC and how early in tumor development gene expression changes in blood cells can be detected. Successful chemopreventive therapy will depend on the elucidation of the network of signaling pathways that regulate the

systemic immune response to the development and presence of a specific tumor, which has its own unique set of genetic, epigenetic and inflammatory changes that evolve with the advancement of disease. Use of our blood-based gene signature for screening of BC now require further improvement for better distinguishing benign breast disease from BC and further compared it with standard mammographic screening. Analyses of gene expression profiles in the matched breast tissue, as well as in blood samples collected within 5 years prior diagnosis, including patients with atypical and *in situ* breast abnormalities have started and hopefully would clarify some of those questions.

Conclusions

In conclusion, gene expression of peripheral blood cells is markedly perturbed by the specific presence of carcinoma in the breast and these changes simultaneously engage a number of systemic cytostatic signals emerging connections with immune escape of BC. Further mining of the cancer-associated blood transcriptome in humans will likely identify additional regulators, mediators and biomarkers of the evolving tumor, its microenvironment and the systemic response to BC and will refine its utility for early detection and treatment of the disease.

Acknowledgements

The authors thank M.T. Hallett for his comments on the data analysis and the article. They also acknowledge S. Cory and R. Lesurf for their assistance in producing the figures, M. Melhus and B. Augdal for the administration of the data and S. Dahl, T. Sauer, T. Cappelen, B. Naume and R. Mortensen member of the Norwegian Breast Cancer Group (NBCG) for their participation in collecting blood samples.

References

1. Bissell MJ, Radisky D. Putting tumours in context. *Nat Rev Cancer* 2001;1:46–54.
2. Cichon MA, Degnim AC, Visscher DW, et al. Microenvironmental influences that drive progression from benign breast disease to invasive breast cancer. *J Mammary Gland Biol Neoplasia* 2010;15:389–97.
3. de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. *Nat Rev Cancer* 2006;6:24–37.
4. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell* 2012;21:309–22.
5. McAllister SS, Gifford AM, Greiner AL, et al. Systemic endocrine instigation of indolent tumor growth requires osteopontin. *Cell* 2008;133:994–1005.
6. Zuckerman NS, Yu H, Simons DL, et al. Altered local and systemic immune profiles underlie lymph node metastasis in breast cancer patients. *Int J Cancer* 2013;132:2537–47.
7. McAllister SS, Weinberg RA. Tumor–host interactions: a far-reaching relationship. *J Clin Oncol* 2010;28:4022–8.
8. Ogawa M. Differentiation and proliferation of hematopoietic stem cells. *Blood* 1993;81:2844–53.
9. Aaroe J, Lindahl T, Dumeaux V, et al. Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Res* 2010;12:R7.
10. Burczynski ME, Twine NC, Dukart G, et al. Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. *Clin Cancer Res* 2005;11:1181–9.
11. Han M, Liew CT, Zhang HW, et al. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. *Clin Cancer Res* 2008;14:455–60.
12. Dumeaux V, Olsen KS, Nuel G, et al. Deciphering normal blood gene expression variation—the NOWAC postgenome study. *PLoS Genet* 2010;6:e1000873.
13. Whitney AR, Diehn M, Popper SJ, et al. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci USA* 2003;100:1896–901.
14. Dumeaux V, Lund E, Borresen-Dale AL. Comparison of globin RNA processing methods for genome-wide transcriptome analysis from whole blood. *Biomark Med* 2008;2:11–21.
15. Debey S, Schoenbeck U, Hellmich M, et al. Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *Pharmacogenomics J* 2004;4:193–207.
16. Dumeaux V, Borresen-Dale AL, Frantzen JO, et al. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Res* 2008;10:R13.
17. Lund E, Kumle M, Braaten T, et al. External validity in a population-based national prospective study—the Norwegian Women and Cancer Study (NOWAC). *Cancer Causes Control* 2003;14:1001–8.
18. Lund E, Dumeaux V, Braaten T, et al. Cohort profile: The Norwegian Women and Cancer Study—NOWAC—Kvinner og kreft. *Int J Epidemiol* 2008;37:36–41.
19. Du P, Kibbe WA, Lin SM. Lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008;24:1547–8.
20. Lin SM, Du P, Huber W, et al. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* 2008;36:e11.
21. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 2010;38:e17.
22. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;3:Article3.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;57:289–300.
24. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? *Int Stat Rev* 2001;69:385–98.
25. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002;97:10.
26. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313:1929–35.
27. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res* 2011;39:D1005–10.
28. LaBreche HG, Nevins JR, Huang E. Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. *BMC Med Genomics* 2011;4:61.
29. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57.
30. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005, 102:15545–50.
31. Birnbaum KD, Kussell E. Measuring cell identity in noisy biological systems. *Nucleic Acids Res* 2011;39:9093–107.
32. Watkins NA, Gusnanto A, de Bono B, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* 2009;113:e1–9.
33. Goeman JJ, van de Geer SA, de Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;20:93–99.
34. Goeman JJ, Finos L. The inheritance procedure: multiple testing of tree-structured hypotheses. *Stat Appl Genet Mol Biol* 2012;11:Article 11.
35. Taler M, Gil-Ad I, Lomnitski L, et al. Immunomodulatory effect of selective serotonin reuptake inhibitors (SSRIs) on human T lymphocyte function and gene expression. *Eur Neuropsychopharmacol* 2007;17:774–80.
36. Gasser R. Myocardial ischemia and the immune system: some thoughts and changing views *J Clin Basic Cardiol* 2011;14:7.
37. Zeller KI, Jegga AG, Aronow BJ, et al. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol* 2003;4:R69.

38. Nie Z, Hu G, Wei G, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* 2012;151:68–79.
39. Shaffer AL, Wright G, Yang L, et al. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. *Immunol Rev* 2006;210:67–85.
40. Neefjes J, Jongsma ML, Paul P, et al. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 2011;11:823–36.
41. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* 2011;331:1565–70.
42. Imai K, Matsuyama S, Miyake S, et al. Natural cytotoxic activity of peripheral-blood lymphocytes and cancer incidence: an 11-year follow-up study of a general population. *Lancet* 2000;356:1795–9.
43. Pitteri SJ, Kelly-Spratt KS, Gurley KE, et al. Tumor microenvironment-derived proteins dominate the plasma proteome response during breast cancer induction and progression. *Cancer Res* 2011;71:5090–100.
44. Warner JR, McIntosh KB. How common are extraribosomal functions of ribosomal proteins? *Mol Cell* 2009;34:3–11.
45. Campoli M, Ferrone S. HLA antigen and NK cell activating ligand expression in malignant cells: a story of loss or acquisition. *Semin Immunopathol* 2011;33:321–34.
46. Croci DO, Zacarias Fluck MF, Rico MJ, et al. Dynamic cross-talk between tumor and immune cells in orchestrating the immunosuppressive network at the tumor microenvironment. *Cancer Immunol Immunother* 2007;56:1687–700.
47. Tavares-Murta BM, Mendonca MA, Duarte NL, et al. Systemic leukocyte alterations are associated with invasive uterine cervical cancer. *Int J Gynecol Cancer* 2010;20:1154–9.
48. Walsh SR, Cook EJ, Goulder F, et al. Neutrophil–lymphocyte ratio as a prognostic factor in colorectal cancer. *J Surg Oncol* 2005;91:181–4.