Database tool

bc-GenExMiner 3.0: new mining module computes breast cancer gene expression correlation analyses

Pascal Jézéquel[1,2,*], Jean-Sébastien Frénel[3], Loïc Campion[2,4], Catherine Guérin-Charbonnel[1,2,4], Wilfried Gouraud[1,2,4], Gabriel Ricolleau[5] and Mario Campone[2,3]

[1]Unité Mixte de Génomique du Cancer, Hôpital Laënnec/Institut de Cancérologie de l'Ouest - site René Gauducheau, Bd J. Monod, 44805 Nantes - Saint Herblain Cedex, [2]INSERM U892, IRT-UN, 8 quai Moncousu, 44007 Nantes Cedex, [3]Service d'Oncologie Médicale, Institut de Cancérologie de l'Ouest - site René Gauducheau, Bd J. Monod, 44805 Nantes - Saint Herblain Cedex, [4]Unité de Biostatistique, Institut de Cancérologie de l'Ouest - site René Gauducheau, Bd J. Monod, 44805 Nantes - Saint Herblain Cedex and [5]Département de Biopathologie, Institut de Cancérologie de l'Ouest - site René Gauducheau, Bd J. Monod, 44805 Nantes - Saint Herblain Cedex, France

***Corresponding author:** Tel: +33 2 40 67 99 00; Fax: +33 2 40 67 97 62; Email: pascal.jezequel@ico.unicancer.fr

Submitted 23 August 2012; Revised 10 December 2012; Accepted 14 December 2012

Citation details: Jézéquel,P., Frénel,J.-S., Campion,L., *et al.* bc-GenExMiner 3.0: new mining module computes breast cancer gene-expression correlation analyses. *Database* (2012) Vol. 2012: article ID bas060; doi:10.1093/database/bas060.

We recently developed a user-friendly web-based application called bc-GenExMiner (http://bcgenex.centregauducheau.fr), which offered the possibility to evaluate prognostic informativity of genes in breast cancer by means of a 'prognostic module'. In this study, we develop a new module called 'correlation module', which includes three kinds of gene expression correlation analyses. The first one computes correlation coefficient between 2 or more (up to 10) chosen genes. The second one produces two lists of genes that are most correlated (positively and negatively) to a 'tested' gene. A gene ontology (GO) mining function is also proposed to explore GO 'biological process', 'molecular function' and 'cellular component' terms enrichment for the output lists of most correlated genes. The third one explores gene expression correlation between the 15 telomeric and 15 centromeric genes surrounding a 'tested' gene. These correlation analyses can be performed in different groups of patients: all patients (without any subtyping), in molecular subtypes (basal-like, HER2+, luminal A and luminal B) and according to oestrogen receptor status. Validation tests based on published data showed that these automatized analyses lead to results consistent with studies' conclusions. In brief, this new module has been developed to help basic researchers explore molecular mechanisms of breast cancer.

Database URL: http://bcgenex.centregauducheau.fr

Introduction

The increasing amount of genomic data represents a new resource for fundamental and translational research, but it is limited in its use due to complexity and heterogeneity of the different studies; therefore, in its raw form, it still remains underexploited. To fully take benefit from this resource, bioinformatics processes, which preserve biological sense caught in annotated genomic data, have to be applied before developing automatized mining functionalities, e.g. biostatistics' analyses. We have recently developed a user-friendly web-based application called bc-GenExMiner, which offered the possibility to evaluate prognostic informativity of genes in breast cancer by means of a 'prognostic module' including three functionalities (1). Statistical analyses were based on genomic data and corresponding bioclinical annotations of 21 studies. In this study, numerous biological tests demonstrated that biological sense contained in breast cancer tumours was preserved despite data origin diversity and bioinformatics process complexity, even when data were merged in new cohorts. This development confirmed our opinion that such data and automatized statistical tests could actually help researchers to find new prognostic markers or therapeutic targets. Hence, we have developed a new module called 'correlation module', which includes three kinds of gene expression correlation analyses. The first one computes correlation coefficient between 2 or more (up to 10) chosen genes. The second one produces two lists of genes that are most correlated (positively and negatively) to a 'tested' gene. A gene ontology (GO) mining function is also proposed to explore GO 'biological process', 'molecular function' and 'cellular component' terms enrichment for the output lists of most correlated genes (2). The third one explores gene expression correlation between a 'tested' gene and each of the 15 DNA 5′- and 15 3′-closest genes surrounding it. The aim of the last functionality is to identify DNA continuous clusters of correlated co-expressed genes, which can be linked to genomic anomalies, including chromosomal aberrations [e.g., copy number alterations (CNAs)]. These correlation analyses can be performed in different groups of patients: all patients (without any subtyping), in molecular subtypes (basal-like, HER2+, luminal A and luminal B) and according to oestrogen receptor (ER) status (3, 4). The interest of testing these subgroups of patients is based on the fact that CNAs differentially affect molecular subtypes (4–6). Validation tests based on published data showed that these automatized analyses lead to results consistent with studies' conclusions. In brief, this new module has been developed to help basic researchers explore molecular mechanisms of breast cancer.

Materials and methods

System architecture and database content

System implementation, data selection and data preprocessing are fully described elsewhere (1).

Briefly, bc-GenExMiner is powered by Apache with a MySQL relational database storage. Web interfaces are written in PHP v5 and JavaScript. Statistical analyses are performed with R software. Datasets included in the database were publicly available (Gene Expression Omnibus, ArrayExpress, Stanford microarray database and also on author's individual web pages). Non-Affymetrix platform data were ratio-normalized, and Affymetrix raw CEL data were MAS5-normalized. Data were then log2-transformed. Finally, to merge data of all studies and create pooled cohorts, data were converted to a common scale (median = 0 and standard deviation = 1).

bc-GenExMiner functionalities

A flowchart details purpose of analyses (Figure 1).

ER status genomic determination

In cohorts GSE1456, GSE3143 and GSE11121, ER status determined by immunohistochemistry was not available.

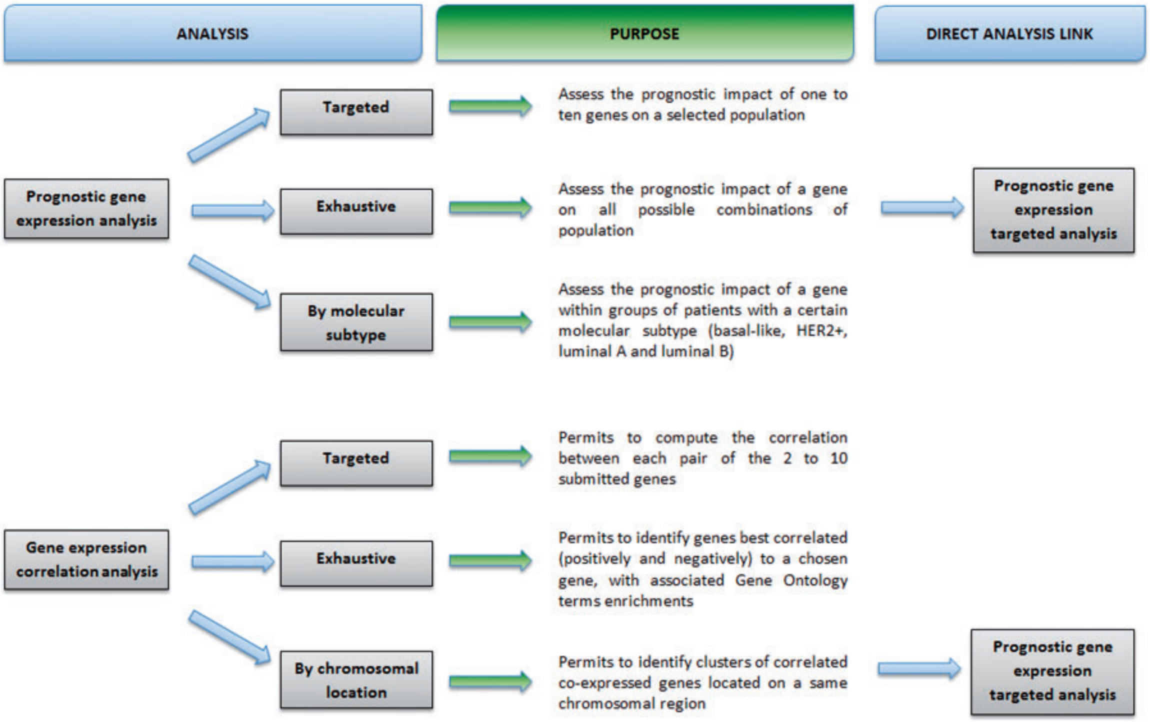

ANALYSIS
PURPOSE
DIRECT ANALYSIS LINK
Prognostic gene expression analysis
Targeted
Exhaustive
By molecular subtype
Assess the prognostic impact of one to ten genes on a selected population
Assess the prognostic impact of a gene on all possible combinations of population
Assess the prognostic impact of a gene within groups of patients with a certain molecular subtype (basal-like, HER2+, luminal A and luminal B)
Prognostic gene expression targeted analysis
Gene expression correlation analysis
Targeted
Exhaustive
By chromosomal location
Permits to compute the correlation between each pair of the 2 to 10 submitted genes
Permits to identify genes best correlated (positively and negatively) to a chosen gene, with associated Gene Ontology terms enrichments
Permits to identify clusters of correlated co-expressed genes located on a same chromosomal region
Prognostic gene expression targeted analysis

Figure 1. bc-GenExMiner 3.0 flowchart.

For these Affymetrix® cohorts, ER status was determined based on 205225_at Affymetrix® probes (U133 array; cohorts GSE1456 and GSE11121) or on the median values of Affymetrix® probes representing *ESR1* (U95 array; cohort GSE3143) using a two-component Gaussian mixture distribution model (3).

Gene expression correlation analyses

Gene correlation targeted analysis. In a gene correlation analysis, the user chooses 2 to 10 different genes, and all possible pairwise Pearson's correlation coefficients are computed, along with the associated *P*-value, on merged datasets based on seven different populations: all patients pooled together or subsets of patients by ER status (two subsets) and molecular subtype (four subsets). Molecular subtype annotation was performed by means of six molecular subtype predictors (MSPs) and lead to robust molecular subtype predictor classification (RMSPC), which only includes patients with concordant molecular subtype assignment for the 6 MSPs (1). Results are displayed in a correlation map, i.e. a table where each cell corresponds to a pairwise correlation and is coloured according to the correlation coefficient value, from dark blue (coefficient $= -1$) to dark red (coefficient $= 1$). Correlation plots are also drawn to illustrate each pairwise correlation.

Gene correlation exhaustive analysis. A gene correlation exhaustive analysis permits to know the genes that are best correlated, positively and negatively, to a given gene, chosen by the user: Pearson's correlation coefficient is computed between the chosen gene and all other genes that are present in bc-GenExMiner database. Analyses are computed on merged datasets and are based on five different sets of patients: all patients, RMSPC basal-like patients, RMSPC HER2+ patients, RMSPC luminal A patients and RMSPC luminal B patients. Genes with coefficient above 0.40 in absolute value and with significant associated *P*-value (<0.05) are retained.

As a complement, GO terms directly linked to genes are screened to identify those underrepresented or overrepresented in the lists of genes that are most positively correlated to the chosen gene (including itself), most negatively correlated to the chosen gene or in the union of the two previous lists.

Results are presented in two tables: one for the positive correlations and one for the negative correlations. For each table, only the 50 best genes are displayed; the full lists can be further downloaded. A GO button permits to visualize GO 'biological process', 'molecular function' and 'cellular component' terms significantly associated with the gene lists. GO analyses are based on the complete lists of genes, which are significantly correlated ($r \geq 0.4$; $P < 0.05$) with the tested gene.

Gene ontology analysis. This functionality is based on the mining of the three GO trees (biological process, molecular function and cellular component) (2). A GO analysis finds GO terms that are significantly linked to a given list of genes ('target list'). For each term of each of the GO trees, comparison is done between the number of occurrences of this term in the 'target list', i.e. the number of times this term is directly linked to a gene, and the number of occurrences of this term in the 'gene universe' (all of the genes that are expressed in the database) by means of Fisher's exact test ($P < 0.01$ is considered significant for this analysis). GO terms database was downloaded from geneontology.org and will be updated every 6 months as gene annotations.

Gene correlation analysis by chromosomal location. DNA location data were computed with information extracted from The Ensembl Genome Database (http://www.ensembl.org), NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/) and NCBI Unigene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) websites. Only gene position (start and end) and DNA strand (positive or negative) were kept in our final list of chromosomal locations.

In a gene correlation analysis by chromosomal location, the user chooses one gene, and Pearson's correlation coefficients, with associated *P*-values, are computed on merged datasets between this gene and genes continuously located around it on the same chromosome (15 centromeric genes and $\leq$15 telomeric genes), based on all patients or on different subsets depending on the ER status or on the molecular subtype. Results are displayed in a table giving the gene location and correlation details for each of the seven different populations (all patients, by ER status, by molecular subtype). Correlation plots are also performed.

Targeted correlation analysis (TCA; 'TCA' button), which here aims at evaluating the robustness of clusters, is proposed: Correlation analyses are automatically computed between all possible pairs of genes that compose a selected cluster.

Biological validation

Targeted analysis. ***MKI67, AURKA*** **and** ***UBE2C:*** We focused on proliferation genes because of the importance of this process on breast cancer prognosis (7). Genomic studies demonstrated the existence of a proliferation cluster containing numerous correlated genes. We chose these three genes because they are known to play an active role in proliferation process in breast cancer: *MKI67* is coding for KI67 protein, which is routinely explored by means of immunohistochemistry, *AURKA* is considered as the proliferation prototypic gene and *UBE2C* bears a high prognostic informativity (8, 9).

ESR1*, *GATA3*, *FOXA1* and *XBP1: Numerous studies, notably based on microarrays, have shown that expressions of *GATA3*, *FOXA1* and *XBP1* were strongly correlated to that of *ESR1* (10).

TNFAIP1*/*POLDIP2*, *RAF1*/*MKRN2* and *TBCB*/*POLR2I: The examples of *TNFAIP1*/*POLDIP2*, *RAF1*/*MKRN2* and *TBCB*/*POLR2I* are of particular interest because these couples of genes demonstrated co-regulatory pattern in breast cancer tumours, are located at the same locus and are organized in sense–antisense architecture on the opposite DNA strands of chromosome 17, 3 and 19, respectively (11).

Exhaustive analysis. ***AURKA***: We chose *AURKA*, which is known to be the prototypic gene of proliferation process. After exhaustive gene expression correlation analysis, we verified that the 72 genes included in proliferation gene expression signature were present in *AURKA*'s most correlated genes list (12).

ESR1: *ESR1*, one of the most important genes in breast cancer physiopathology, was tested.

FTL: We recently validated ferritin light chain (FTL) as a breast cancer prognostic marker (13). In this study, we demonstrated by means of double immunofluorescence that FTL was located in tumour-associated macrophages (TAMs) harbouring an M2-phenotype (CD163-positive TAM).

Gene ontology analysis. Following gene correlation exhaustive analysis, GO biological process enrichment analyses for *AURKA*, *ESR1* and *FTL* were computed to explore biological process annotations of lists of most correlated genes.

By chromosomal location. This kind of analysis pinpoints continuous clusters of co-expressed genes and permits to visualize their chromosomal organization. For all following tests, TCA was performed to verify the robustness of clusters.

An *in silico* study conducted by Buness *et al*. (14) identified 32 series of 20 characteristic overexpressed genes for amplified chromosomal regions. Six hundred and forty genes were tested for all patients and molecular subtypes. We considered as positive results continuous clusters of correlated co-expressed genes composed of at least three genes with $r \geq 0.30$ and $P \leq 0.05$. To quantify the global relationship among a cluster of co-expressed genes, eigenvalues of the correlation matrix were computed, and the ratio of the largest one to the sum of all eigenvalues multiply by 100 was taken (i.e. it comes to perform a principal component analysis based on the correlation matrix and take the percentage of variance explained by the first principal component) (Supplementary Method). This value is called multicorrelation score (MCS).

ESR1: A recent study showed that *ESR1* was co-expressed with closely adjacent genes at 6q25.1 (15).

8p11-12 amplicon: At 8p11-12, one of the most frequently amplified regions in breast cancer (10–15% of cases), Bernard-Pierrot *et al*. identified five genes (*LSM1*, *BAG4*, *DDHD2*, *PPAPDC1B* and *WHSC1L1*) as consistently overexpressed due to an increased gene copy number (16, 17).

Chromosome 17: Chromosome 17 is highly amplified in breast cancer, especially in HER2+ molecular subtype. Its amplified regions permit to distinguish HER2+ and luminal A (5, 18). We chose three genes in different regions of chromosome 17 to test bc-GenExMiner analyses: *TRAF4* at 17q11-q12, *MED24* at 17q21.1 and *GGA3* at 17q25.1 (14).

ER status: Numerous studies showed that CNA varied among breast tumours with different ER status (4, 5, 19–21). A list of 59 genes, which demonstrated different expression profile according to ER status in gained regions, was tested to screen for DNA continuous clusters of correlated co-expressed genes (20). We chose the following criterion: at least two DNA continuous genes with correlated gene-expressions, to define clusters of correlated co-expressed genes. MCS was used to compare clusters of co-expressed genes.

Differential gene expression according to ER status. As a complement of correlation analyses by chromosomal location within homogeneous ER status groups, expression of genes according to ER status was also studied by means of Mann–Whitney test.

Remark: As *ESR1*, *ERBB2* and *AURKA* are the three genes most implicated in the molecular subtypes determination (*ESR1*, *ERBB2* and *AURKA* play indeed a very important role in the determination of the molecular subtypes whatever the MSP chosen; in particular, one of the six predictors involved in the RMSPC only takes those three genes into account to determine a patient's molecular subtype), they were not tested within each of those subtypes to avoid bias in the analysis results due to the dependency between molecular subtype determination and gene expression.

Results

Biological validation

Correlation analyses. **Targeted**: M*KI67*, *AURKA* and *UBE2C* demonstrated significant correlations in 'all patients': $r = 0.74$ for *AURKA*/*UBE2C* (n = 2928), 0.64 for *AURKA*/*MKI67* ($n = 2928$) and 0.58 *MKI67*/*UBE2C* ($n = 3160$) with $P < 10^{-4}$.

***ESR1*, *GATA3*, *FOXA1* and *XBP1*:** As suspected, bc-GenExMiner analysis showed that all these genes were correlated, with correlation coefficients ranging from 0.51 to 0.73 (Table 1).

As expected, *TNFAIP1/POLDIP2, RAF1/MKRN2 and TBCB/POLR2I* correlation analyses showed correlated

Table 1. Targeted correlation analysis of *ESR1*, *GATA3*, *FOXA1* and *XBP1* ($P < 10^{-4}$)

	GATA3	*FOXA1*	*XBP1*
ESR1	$r = 0.56$ ($n = 3414$)	$r = 0.52$ ($n = 3315$)	$r = 0.51$ ($n = 3404$)
GATA3	—	$r = 0.73$ ($n = 3315$)	$r = 0.60$ ($n = 3404$)
FOXA1	—	—	$r = 0.66$ ($n = 3315$)

Table 2. Results of gene correlation exhaustive analysis and GO enrichment analysis for the biological process tree for *AURKA*, *ESR1* and *FTL*, 'all patients' population

Tested genes	No. positively correlated genes[a]	No. genes with GO biological process annotations[a]	No. GO enrichment terms ($P < 0.01$)
AURKA	590	365	121
ESR1	487	215	38
FTL	281	172	78

[a] Includes reference gene.

co-expressions for the three couples of genes ($r = 0.51$, 0.44 and 0.56, respectively, with $P < 10^{-4}$).

Exhaustive: Summary of results from exhaustive correlation analyses for *AURKA*, *ESR1* and *FTL* is displayed in Table 2.

AURKA: Each gene belonging to the 72-gene expression proliferative signature was present in *AURKA* 'correlation exhaustive analysis' output, which meant each correlation coefficient was superior to 0.40 and significant (actually, all *P*-values were $< 10^{-4}$).

FTL: Following one of our recent work, *FTL* and *CD163* gene expression were checked (13). *FTL* correlation exhaustive analysis output showed that *CD163* gene was correlated with *FTL* gene in breast cancer groups of patients with or without molecular subtyping ($r = 0.44$–0.54). In a cohort of 2773 patients, *CD163* was the 15th most correlated with *FTL* gene (Table 3). Rank varied according to molecular subtype.

Gene ontology analysis: The results of GO biological process enrichment term analyses for *AURKA*, *ESR1* and *FTL* are displayed in Table 2.

AURKA: As suspected, GO enrichment for biological process of *AURKA*'s most correlated genes essentially pointed out proliferation via the following terms: 'mitosis', 'cell division', 'cell cycle' and 'DNA replication'.

ESR1: Mammary development, represented by *AR*, *PGR*, *ESR1* and *CCND1* genes, and oestrogen pathway, represented by *EGLN2*, *RARA*, *ESR1*, *GATA3*, *BCL2*, *CRIPACK*, *CCND1*, *FOXA1* and *AR* genes, appeared in the best classified biological process terms. Prostate and male gonad development also appeared in the top of this classification because genes involved in these processes also play a major role in mammary molecular physiology (*FOXA1*, *AR* and *IGF1R* for 'prostate gland epithelium morphogenesis', *BBS2* and *UBE2B* for 'sperm axoneme assembly' and *ESR1*, *GATA3*, *BCL2*, *AR* and *PATZ1* for 'male gonad development').

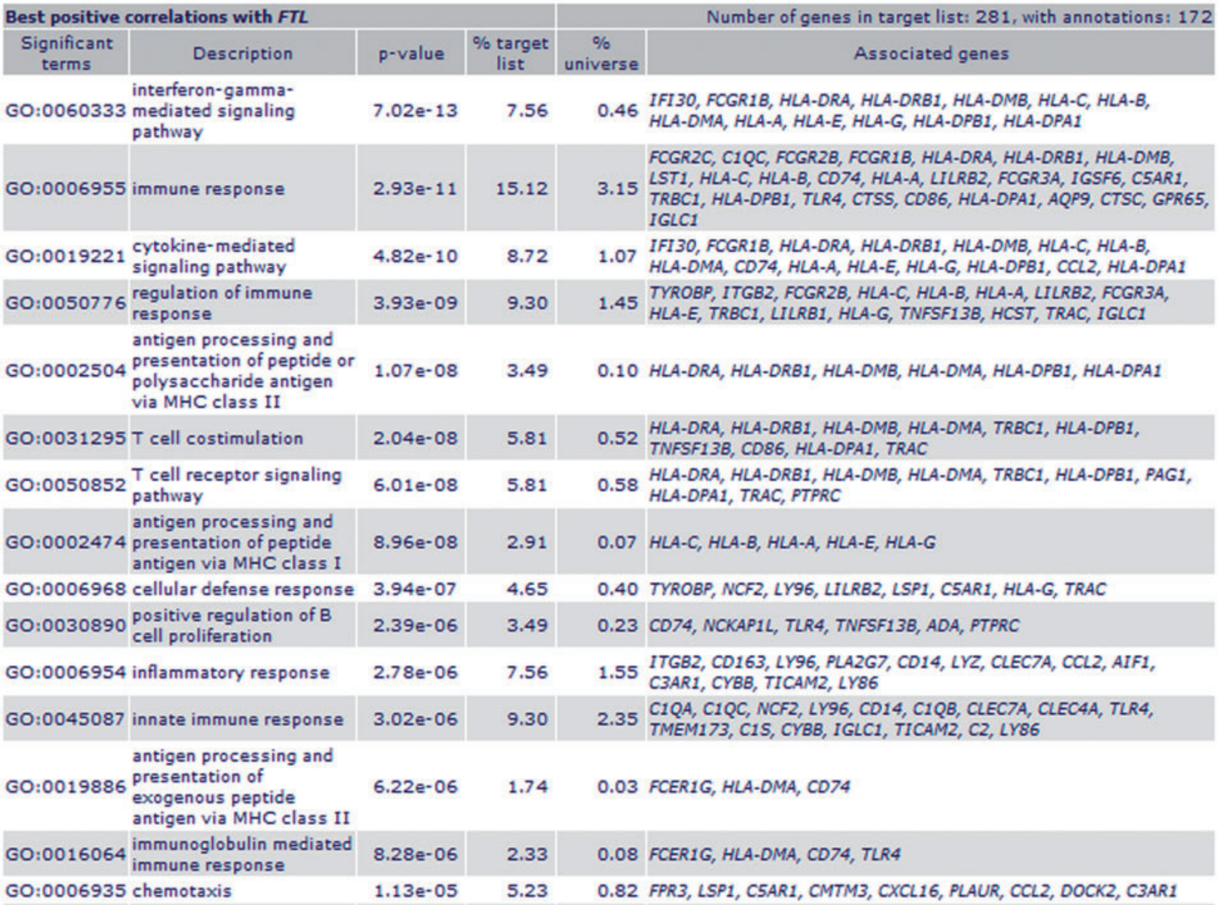
Table 3. Summary results of *FTL* gene correlation exhaustive analysis, with focus on *CD163*

Breast cancer patients	No. of patients	No. *FTL* positively correlated genes ($r > 0.4$ and $P < 0.05$)	Rank of *CD163*	*CD163/FTL* r ($P < 0.001$)
All patients	2773	280	15	0.54
Basal-like	274	107	38	0.49
HER2+	76	569	134	0.54
Luminal A	194	190	118	0.44
Luminal B	53	286	170	0.44

FTL: GO enrichment for biological process of *FTL*'s most correlated genes pointed out immune response, immune cells (T and B cells, neutrophils and macrophages), immune processes and inflammatory response (Figure 2). 'Positive regulation of macrophage chemotaxis' included *C5AR1*, *C3AR1* and *RARRES2* genes.

By chromosomal location: On the basis of Buness data, we always found continuous clusters of correlated co-expressed genes for all 32 chromosomal regions (Supplementary Table 1). The largest cluster, located at 17q24-q25 cytoband, was composed of 12 genes (*CDR2L*, *ICT1*, *ATP5H*, *KCTD2*, *SLC16A5*, *ARMC7*, *NT5C*, *HN1*, *SUMO2*, *NUP85*, *GGA3* and *MRPS7*) for HER2 molecular subtype. Other large clusters were identified in 17q cytobands for HER2 molecular subtype (17q12-q21, 17q11-q12 and 17q21-q23). Consistent with numerous studies, 17q-clusters were rarely found in luminal A molecular subtype.

ESR1: Among the 30 genes selected in the close vicinity of *ESR1* (*PCMT1 - LRP11 - RAET1E - ULBP2 - ULBP1 - ULBP3 - PPP1R14C - IYD - PLEKHG1 - MTHFD1L - AKAP12 - ZBTB2 RMND1 - C6ORF211 - C6ORF97 - [ESR1] - SYNE1 - MYCT1 – VIP - FBXO5 - MTRF1L - RGS17 - OPRM1 - IPCEF1 - CNKSR3 - RBM16 - TIAM2 - CLDN20 - TFB1M - NOX3 - ARID1B*), only 3 genes, all located in a continuous amplicon containing ER gene at 6q25.1, presented a good correlation with *ESR1* in all patients with breast cancer (without molecular subtyping): *C6orf211*, *C6orf97* and *RMND1* (*C6orf96*), even when correlation targeted analysis including these 4 genes was performed. Furthermore, level of expression of these genes according to ER status, based on bcGenExMiner data, showed that all of them were overexpressed in ER+



Best positive correlations with *FTL*				Number of genes in target list: 281, with annotations: 172	
Significant terms	Description	p-value	% target list	% universe	Associated genes
GO:0060333	interferon-gamma-mediated signaling pathway	7.02e-13	7.56	0.46	*IFI30, FCGR1B, HLA-DRA, HLA-DRB1, HLA-DMB, HLA-C, HLA-B, HLA-DMA, HLA-A, HLA-E, HLA-G, HLA-DPB1, HLA-DPA1*
GO:0006955	immune response	2.93e-11	15.12	3.15	*FCGR2C, C1QC, FCGR2B, FCGR1B, HLA-DRA, HLA-DRB1, HLA-DMB, LST1, HLA-C, HLA-B, CD74, HLA-A, LILRB2, FCGR3A, IGSF6, C5AR1, TRBC1, HLA-DPB1, TLR4, CTSS, CD86, HLA-DPA1, AQP9, CTSC, GPR65, IGLC1*
GO:0019221	cytokine-mediated signaling pathway	4.82e-10	8.72	1.07	*IFI30, FCGR1B, HLA-DRA, HLA-DRB1, HLA-DMB, HLA-C, HLA-B, HLA-DMA, CD74, HLA-A, HLA-E, HLA-G, HLA-DPB1, CCL2, HLA-DPA1*
GO:0050776	regulation of immune response	3.93e-09	9.30	1.45	*TYROBP, ITGB2, FCGR2B, HLA-C, HLA-B, HLA-A, LILRB2, FCGR3A, HLA-E, TRBC1, LILRB1, HLA-G, TNFSF13B, HCST, TRAC, IGLC1*
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.07e-08	3.49	0.10	*HLA-DRA, HLA-DRB1, HLA-DMB, HLA-DMA, HLA-DPB1, HLA-DPA1*
GO:0031295	T cell costimulation	2.04e-08	5.81	0.52	*HLA-DRA, HLA-DRB1, HLA-DMB, HLA-DMA, TRBC1, HLA-DPB1, TNFSF13B, CD86, HLA-DPA1, TRAC*
GO:0050852	T cell receptor signaling pathway	6.01e-08	5.81	0.58	*HLA-DRA, HLA-DRB1, HLA-DMB, HLA-DMA, TRBC1, HLA-DPB1, PAG1, HLA-DPA1, TRAC, PTPRC*
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	8.96e-08	2.91	0.07	*HLA-C, HLA-B, HLA-A, HLA-E, HLA-G*
GO:0006968	cellular defense response	3.94e-07	4.65	0.40	*TYROBP, NCF2, LY96, LILRB2, LSP1, C5AR1, HLA-G, TRAC*
GO:0030890	positive regulation of B cell proliferation	2.39e-06	3.49	0.23	*CD74, NCKAP1L, TLR4, TNFSF13B, ADA, PTPRC*
GO:0006954	inflammatory response	2.78e-06	7.56	1.55	*ITGB2, CD163, LY96, PLA2G7, CD14, LYZ, CLEC7A, CCL2, AIF1, C3AR1, CYBB, TICAM2, LY86*
GO:0045087	innate immune response	3.02e-06	9.30	2.35	*C1QA, C1QC, NCF2, LY96, CD14, C1QB, CLEC7A, CLEC4A, TLR4, TMEM173, C1S, CYBB, IGLC1, TICAM2, C2, LY86*
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	6.22e-06	1.74	0.03	*FCER1G, HLA-DMA, CD74*
GO:0016064	immunoglobulin mediated immune response	8.28e-06	2.33	0.08	*FCER1G, HLA-DMA, CD74, TLR4*
GO:0006935	chemotaxis	1.13e-05	5.23	0.82	*FPR3, LSP1, C5AR1, CMTM3, CXCL16, PLAUR, CCL2, DOCK2, C3AR1*

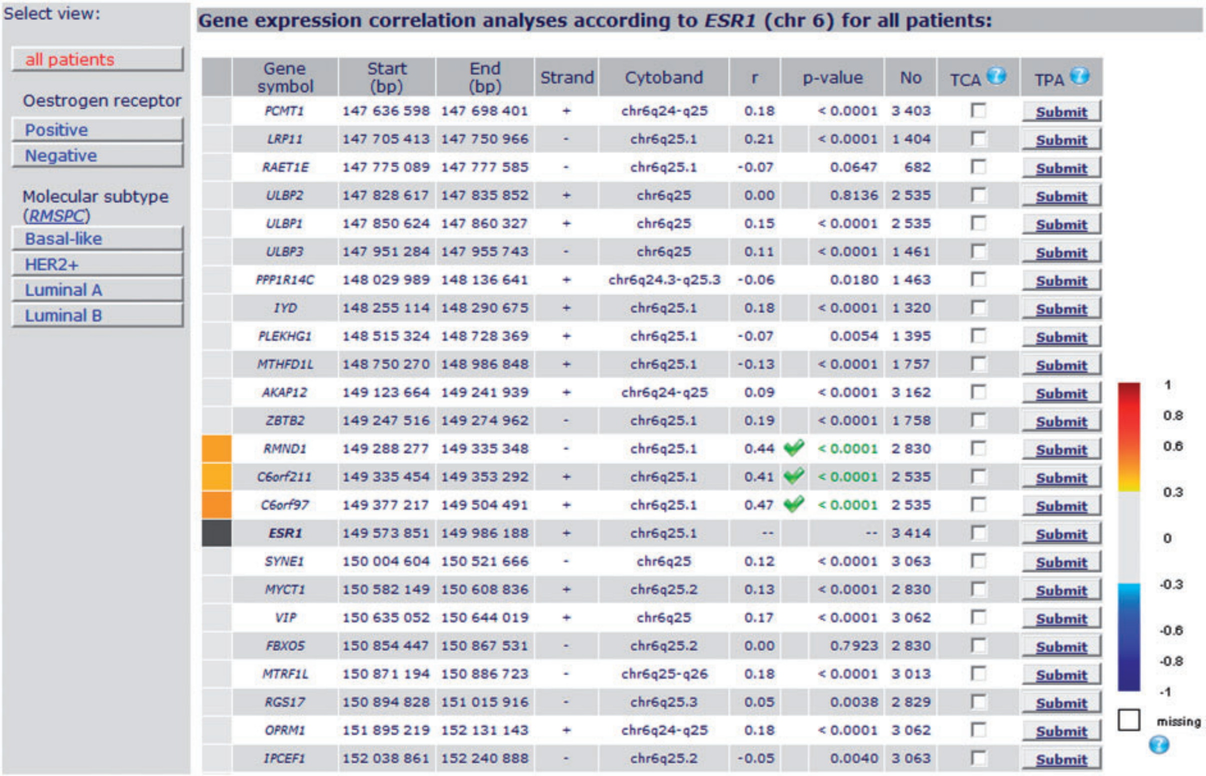
Figure 2. Biological process ontology enrichment results for FTL most positively correlated genes in all patients population (first 15 GO biological process terms).

tumours (Supplementary Figure 1). bcGenExMiner results are concordant with Dunbier's ones (15) (Figure 3).

8p11-12 amplicon: Among the 30 genes continuously located around *LSM1* gene (*DUSP26 - C8ORF41 - UNC5D - KCNU1 - ZNF703 - ERLIN2 - PROSC - GPR124 - BRF2 - RAB11FIP1 - GOT1L1 - ADRB3 - EIF4EBP1 - ASH2L - STAR - [LSM1] - BAG4 - DDHD2 - PPAPDC1B - WHSC1L1 - FKSG2 - LETM2 - FGFR1 - TACC1 - PLEKHA2 - HTRA4 - TM2D2 - ADAM9 - ADAM32 - ADAM5P - ADAM3A*), bc-GenExMiner correlation analysis showed a robust continuous cluster of co-expressed genes composed of *LSM1*, *BAG4*, *DDHD2*, *PPAPDC1B* and *WHSC1L1* in 'all patients' group (Figure 4). This result was in agreement with Bernard-Pierrot's one (16). Robustness of this cluster was demonstrated by performing a TCA with these five genes (Table 4). This cluster was also found in basal-like, HER2+ and luminal B patients. Best MCS (87.7) was obtained in luminal B patients, who display more frequent high-level DNA amplification, especially in chromosome 8 region, than other molecular subtypes (MCS = 63.1 in basal-like and 71.5 in HER2) (5) (Table 4). This last result strengthened the idea that genomic instability, here indirectly materialized by cluster of co-expressed genes, may be specific of molecular subtypes and that bc-GenExMiner may help researchers to find such genomic anomalies.

Chromosome 17: *TRAF4*, *MED24* and *GGA3* are included in clusters of correlated co-expressed genes, composed of 8, 5 and 13 genes, respectively (Table 5). Correlation robustness was tested by means of TCA. These clusters differentiate HER2+ from luminal A subtype, which did not display the same expression profile. In this molecular subtype, there is no correlation between the genes of the above-mentioned clusters.

ER status: Of the 59 tested genes, 21 belonged to continuous clusters of correlated co-expressed genes specific of ER− status and 4 to amplicons of discontinuous correlated co-expressed genes (Supplementary Table 2). No ER− specificity was found for 34 of these genes. In this group, 24 were not linked to ER− or ER+ status while 10 showed ER+ specificity, the last ones are called discordant cases. In discordant cases, seven showed a basal-like specificity or tendency; these contradictory results question the



Select view:

all patients

Oestrogen receptor

Positive

Negative

Molecular subtype (RMSPC)

Basal-like

HER2+

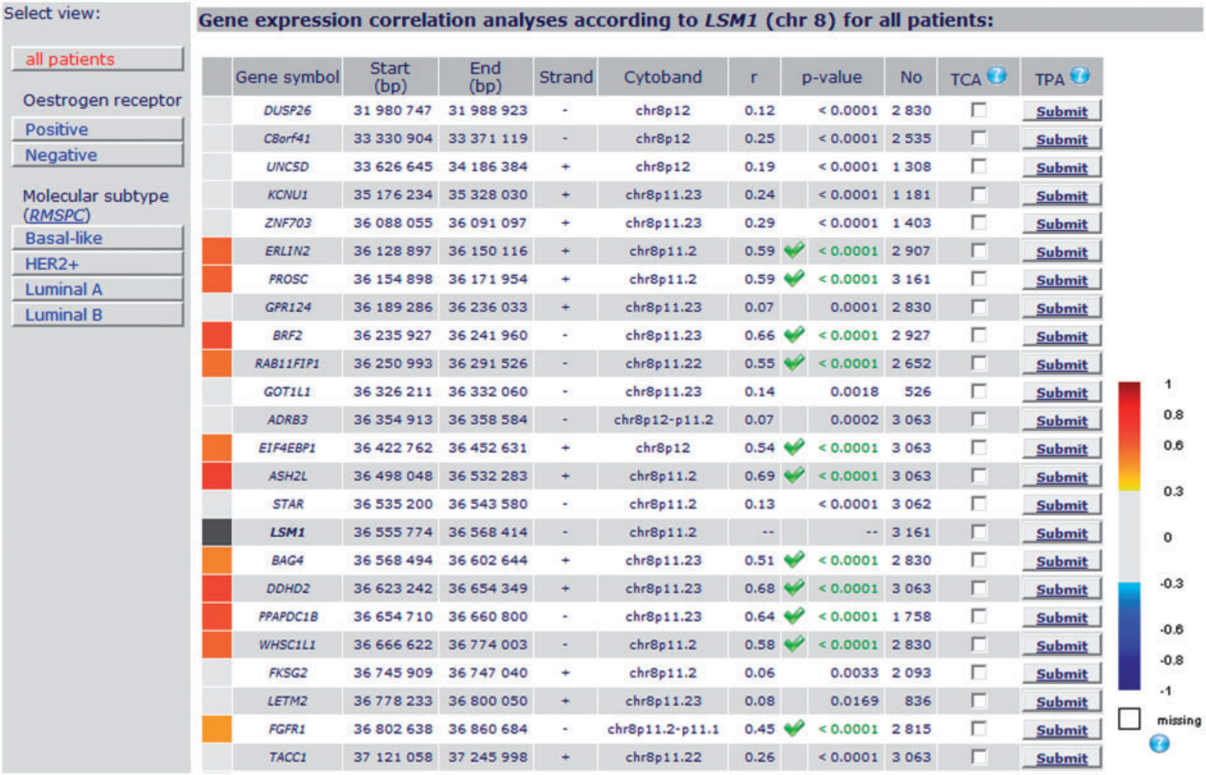
Luminal A

Luminal B

Gene expression correlation analyses according to *ESR1* (chr 6) for all patients:

	Gene symbol	Start (bp)	End (bp)	Strand	Cytoband	r		p-value	No	TCA	TPA
	PCMT1	147 636 598	147 698 401	+	chr6q24-q25	0.18		< 0.0001	3 403	☐	Submit
	LRP11	147 705 413	147 750 966	-	chr6q25.1	0.21		< 0.0001	1 404	☐	Submit
	RAET1E	147 775 089	147 777 585	-	chr6q25.1	-0.07		0.0647	682	☐	Submit
	ULBP2	147 828 617	147 835 852	+	chr6q25	0.00		0.8136	2 535	☐	Submit
	ULBP1	147 850 624	147 860 327	+	chr6q25	0.15		< 0.0001	2 535	☐	Submit
	ULBP3	147 951 284	147 955 743	-	chr6q25	0.11		< 0.0001	1 461	☐	Submit
	PPP1R14C	148 029 989	148 136 641	+	chr6q24.3-q25.3	-0.06		0.0180	1 463	☐	Submit
	IYD	148 255 114	148 290 675	+	chr6q25.1	0.18		< 0.0001	1 320	☐	Submit
	PLEKHG1	148 515 324	148 728 369	+	chr6q25.1	-0.07		0.0054	1 395	☐	Submit
	MTHFD1L	148 750 270	148 986 848	+	chr6q25.1	-0.13		< 0.0001	1 757	☐	Submit
	AKAP12	149 123 664	149 241 939	+	chr6q24-q25	0.09		< 0.0001	3 162	☐	Submit
	ZBTB2	149 247 516	149 274 962	-	chr6q25.1	0.19		< 0.0001	1 758	☐	Submit
🟧	*RMND1*	149 288 277	149 335 348	-	chr6q25.1	0.44	✓	< 0.0001	2 830	☐	Submit
🟧	*C6orf211*	149 335 454	149 353 292	+	chr6q25.1	0.41	✓	< 0.0001	2 535	☐	Submit
🟧	*C6orf97*	149 377 217	149 504 491	+	chr6q25.1	0.47	✓	< 0.0001	2 535	☐	Submit
⬛	***ESR1***	149 573 851	149 986 188	+	chr6q25.1	--		--	3 414	☐	Submit
	SYNE1	150 004 604	150 521 666	-	chr6q25	0.12		< 0.0001	3 063	☐	Submit
	MYCT1	150 582 149	150 608 836	+	chr6q25.2	0.13		< 0.0001	2 830	☐	Submit
	VIP	150 635 052	150 644 019	+	chr6q25	0.17		< 0.0001	3 062	☐	Submit
	FBXO5	150 854 447	150 867 531	-	chr6q25.2	0.00		0.7923	2 830	☐	Submit
	MTRF1L	150 871 194	150 886 723	-	chr6q25-q26	0.18		< 0.0001	3 013	☐	Submit
	RGS17	150 894 828	151 015 916	-	chr6q25.3	0.05		0.0038	2 829	☐	Submit
	OPRM1	151 895 219	152 131 143	+	chr6q24-q25	0.18		< 0.0001	3 062	☐	Submit
	IPCEF1	152 038 861	152 240 888	-	chr6q25.2	-0.05		0.0040	3 063	☐	Submit

Figure 3. Detailed results of *ESR1* gene expression correlation analysis by chromosomal location for all patients.



Select view:

all patients

Oestrogen receptor

Positive

Negative

Molecular subtype (RMSPC)

Basal-like

HER2+

Luminal A

Luminal B

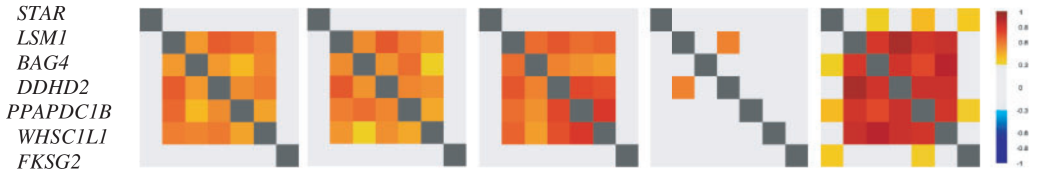
Gene expression correlation analyses according to *LSM1* (chr 8) for all patients:

	Gene symbol	Start (bp)	End (bp)	Strand	Cytoband	r		p-value	No	TCA	TPA
	DUSP26	31 980 747	31 988 923	-	chr8p12	0.12		< 0.0001	2 830	☐	Submit
	C8orf41	33 330 904	33 371 119	-	chr8p12	0.25		< 0.0001	2 535	☐	Submit
	UNC5D	33 626 645	34 186 384	+	chr8p12	0.19		< 0.0001	1 308	☐	Submit
	KCNU1	35 176 234	35 328 030	+	chr8p11.23	0.24		< 0.0001	1 181	☐	Submit
	ZNF703	36 088 055	36 091 097	+	chr8p11.23	0.29		< 0.0001	1 403	☐	Submit
🟧	*ERLIN2*	36 128 897	36 150 116	+	chr8p11.2	0.59	✓	< 0.0001	2 907	☐	Submit
🟧	*PROSC*	36 154 898	36 171 954	+	chr8p11.2	0.59	✓	< 0.0001	3 161	☐	Submit
	GPR124	36 189 286	36 236 033	+	chr8p11.23	0.07		0.0001	2 830	☐	Submit
🟥	*BRF2*	36 235 927	36 241 960	-	chr8p11.23	0.66	✓	< 0.0001	2 927	☐	Submit
🟧	*RAB11FIP1*	36 250 993	36 291 526	-	chr8p11.22	0.55	✓	< 0.0001	2 652	☐	Submit
	GOT1L1	36 326 211	36 332 060	-	chr8p11.23	0.14		0.0018	526	☐	Submit
	ADRB3	36 354 913	36 358 584	-	chr8p12-p11.2	0.07		0.0002	3 063	☐	Submit
🟧	*EIF4EBP1*	36 422 762	36 452 631	+	chr8p12	0.54	✓	< 0.0001	3 063	☐	Submit
🟥	*ASH2L*	36 498 048	36 532 283	+	chr8p11.2	0.69	✓	< 0.0001	3 063	☐	Submit
	STAR	36 535 200	36 543 580	-	chr8p11.2	0.13		< 0.0001	3 062	☐	Submit
⬛	***LSM1***	36 555 774	36 568 414	-	chr8p11.2	--		--	3 161	☐	Submit
🟧	*BAG4*	36 568 494	36 602 644	+	chr8p11.23	0.51	✓	< 0.0001	2 830	☐	Submit
🟥	*DDHD2*	36 623 242	36 654 349	+	chr8p11.23	0.68	✓	< 0.0001	3 063	☐	Submit
🟥	*PPAPDC1B*	36 654 710	36 660 800	-	chr8p11.23	0.64	✓	< 0.0001	1 758	☐	Submit
🟧	*WHSC1L1*	36 666 622	36 774 003	-	chr8p11.2	0.58	✓	< 0.0001	2 830	☐	Submit
	FKSG2	36 745 909	36 747 040	+	chr8p11.2	0.06		0.0033	2 093	☐	Submit
	LETM2	36 778 233	36 800 050	+	chr8p11.23	0.08		0.0169	836	☐	Submit
🟧	*FGFR1*	36 802 638	36 860 684	-	chr8p11.2-p11.1	0.45	✓	< 0.0001	2 815	☐	Submit
	TACC1	37 121 058	37 245 998	+	chr8p11.22	0.26		< 0.0001	3 063	☐	Submit

Figure 4. Detailed results of *LSM1* gene expression correlation analysis by chromosomal location for all patients.

Database, Vol. 2013, Article ID bas060, doi:10.1093/database/bas060

Table 4. Targeted correlation analysis of *LSM1*, *BAG4*, *DDHD2*, *PPAPDC1B* and *WHSC1L1* cluster at 8p11-12 in all breast cancer patients and within molecular subtypes



	All patients	Basal-like	HER2	Luminal A	Luminal B	r
MCS[a] (%)	64.14	63.11	71.50	40.72	86.70	

[a]Multicorrelation score for *LSM1* - *BAG4* - *DDHD2* - *PPAPDC1B* - *WHSC1L1* cluster.

Table 5. Robust and continuous correlated co-expressed gene clusters of chromosome 17

Tested gene	Cytoband	Genes of the cluster	Length of non-amplified DNA region covered by the cluster (bp)
TRAF4	17q11-q12	*TLCD1*, *NEK8*, *TRAF4*, *C17orf63*, *ERAL1*, *FLOT2*, *DHRS13*, *PHF12*	226 406
MED24	17q21.1	*MED24*, *THRA*, *NR1D1*, *MSL1*, *CASC3*	152 273
GGA3	17q25.1	*C17orf28*, *CDR2L*, *ICT1*, *ATP5H*, *KCTD2*, *SLC16A5*, *ARMC7*, *NT5C*, *HN1*, *SUMO2*, *NUP85*, *GGA3*, *MRPS7*	315 466

robustness of ER+ specificity in most of these cases. By means of bc-GenExMiner gene expression correlation analysis by chromosomal location, we found a majority of clusters of correlated co-expressed genes linked to ER− status. Discordant results might be explained by a bias in the study by Han et al. due to low number of cases compared to those used in our analyses.

Discussion

Gene expression correlation analysis permits to explore mathematical relation between two genes; precisely, the strength of the positive or negative linear link. But one must not conclude that correlated genes are involved in the same causal chain of events, even if it might happen sometimes. Causal chain involves correlated parameters, but the reverse is untrue. We have to remind that correlation is different from causal determinism. Causal determinism means a physical link, which is time dependent as defined by the causal sequence: A cause produces an effect. In molecular biology, correlation between gene expressions might more likely deal with co-regulation process. Correlation may be used to detect causal connections, which need further and rigorous experimental investigations to be proved.

Gene expression TCA permits to explore link between pairs of genes in an intuitive manner. Therefore, the user can easily and rapidly test its intuition or benchmark results.

Gene correlation exhaustive analysis extracts genes that are correlated with the gene of interest in a non-intuitive, i.e. screening, manner. Computation involves all genes included in bc-GenExMiner genomic database ($n = 20306$). Non-intuitive and automatized interpretation by means of GO terms enrichment is proposed to help researchers to give a biological sense to these data.

Gene correlation analysis by chromosomal location has been developed to identify DNA continuous clusters of correlated co-expressed genes in a cohort composed of all patients or on different molecular subtype breast cancer patients, and so to evaluate whether these clusters are specific of molecular subtypes. Co-expression may be due to co-regulation process, which can be linked to DNA amplification, i.e. DNA CNA. Previous studies demonstrated that highly amplified genes (44–62%) showed moderately or highly elevated expression (22, 23). In the opposite way, in 10.5 to 12%, overexpression was directly attributable to variation in gene copy number. DNA amplification may involve one gene or several genes belonging to a same locus. Our hypothesis was that amplified or deleted genes might be co-expressed, i.e. overexpressed or not expressed, due to these genomic mutations and that our analysis might complete DNA CNA screening studies by means of comparative genomic hybridization or single-nucleotide polymorphism arrays.

Supplementary Data

Supplementary data are available at *Database* Online.

Funding

Funding for open access charge: Our institution's (ICO) research budget.

Conflict of interest. None declared.

References

1. Jézéquel,P., Campone,M., Gouraud,W. *et al*. (2012) bc-GenExMiner: an easy-to-use online platform for gene prognostic analyses in breast cancer. *Breast Cancer Res. Treat.*, **131**, 765–774.
2. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
3. Lehmann,B.D., Bauer,J.A., Chen,X. *et al*. (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, **121**, 2750–2767.
4. Loo,L.W., Grove,D.I., Williams,E.M. *et al*. (2004) Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res.*, **64**, 8541–8549.
5. Bergamaschi,A., Kim,Y.H., Wang,P. *et al*. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer*, **45**, 1033–1040.
6. Chin,K., DeVries,S., Fridlyand,J. *et al*. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.
7. Wirapati,P., Sotiriou,C., Kunkel,S. *et al*. (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.*, **10**, R65.
8. Jézéquel,P., Campone,M., Roché,H. *et al*. (2009) 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant chemotherapy: a genomic substudy of PACS01 clinical trial. *Breast Cancer Res. Treat.*, **116**, 509–520.
9. Loussouarn,D., Campion,L., Leclair,F. *et al*. (2009) Validation of UBE2C as a prognostic marker in node-positive breast cancer. *Br. J. Cancer*, **101**, 166–173.
10. Lacroix,M. and Leclercq,G. (2004) About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-α gene (ESR1) in breast cancer. *Mol. Cell Endocrinol.*, **219**, 1–7.
11. Grinchuk,O.V., Motakis,E. and Kuznetsov,V.A. (2010) Complex sense-antisense architecture of TNAIP1/POLDIP2 on 17q11.2 represents a novel transcriptional structural-functional gene module involved in breast cancer progression. *BMC Genomics*, **11**, 1S9.
12. Dexter,T.J., Sims,D., Mitsopoulos,C. *et al*. (2010) Genomic distance entrained clustering and regression modeling highlights interacting genomic regions contributing to proliferation in breast cancer. *BMC Syst. Biol.*, **4**, 127.
13. Jézéquel,P., Campion,L., Spyratos,F. *et al*. (2012) Validation of tumour-associated macrophage ferritin light chain as a prognostic biomarker in node-negative breast cancer tumours: a multicentric 2004 national PHRC study. *Int. J. Cancer*, **131**, 426–437.
14. Buness,A., Kuner,R., Ruschhaupt,M. *et al*. (2007) Identification of aberrant chromosomal regions from gene expression microarray studies applied to human breast cancer. *Bioinformatics*, **23**, 2273–2280.
15. Dunbier,A.K., Anderson,H., Ghazoui,Z. *et al*. (2011) ESR1 is co-expressed with closely adjacent uncharacterized genes spanning a breast cancer susceptibility locus at 6q25.1. *PLoS Genet.*, e1001382.
16. Bernard-Pierrot,I., Gruel,N., Stransky,N. *et al*. (2008) Characterization of the recurrent 8p11-12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer. *Cancer Res.*, **68**, 7165–7175.
17. André,F., Job,B., Dessen,P. *et al*. (2009) Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin. Cancer Res.*, **15**, 441–451.
18. Hu,X., Stern,H.M., Ge,L. *et al*. (2009) Genetic alterations and oncogenic pathways associated with breast cancer subtypes. *Mol. Cancer Res.*, **7**, 511–522.
19. Pierga,J.Y., Reis-Filho,J.S., Cleator,S.J. *et al*. (2007) Microarray-based comparative genomic hybridization of breast cancer patients receiving neoadjuvant chemotherapy. *Br. J. Cancer*, **96**, 341–351.
20. Han,W., Jung,E.M., Cho,J. *et al*. (2008) DNA copy number alterations and expression of relevant genes in triple-negative breast cancer. *Genes Chromosomes Cancer*, **47**, 490–499.
21. Horlings,H.M., Lai,C., Nuyten,D.S.A. *et al*. (2010) Integration of DNA copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clin. Cancer Res.*, **16**, 651–663.
22. Hyman,E., Kaurianemi,P., Hautaniemi,S. *et al*. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.
23. Pollack,J.R., Sorlie,T., Perou,C.M. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.