Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Dogan et al., J Biom Biostat 2016, 7:2
DOI: 10.4172/2155-6180.1000286

Research Article Open Access

# The Detection of Extremely High and Low Expressed Genes by EGEF Algorithm in Invasive Breast Cancer

**Senol Dogan[1]*, Amina Kurtovic-Kozaric[1,2] and Gunay Karlı[3]**

[1]*Genetics and Bioengineering Department, International Burch University, Sarajevo, Bosnia and Herzegovina*
[2]*Clinical Pathology and University Clinical Center, Sarajevo University, Sarajevo, Bosnia and Herzegovina*
[3]*Information Technologies Department, International Burch University, Sarajevo, Bosnia and Herzegovina*

## Abstract

Invasive breast cancer is a heterogeneous disease. The analysis of one or a group of specific gene expression profiles may not be enough to understand molecular activities in cancer cells. Therefore, a method which gives us the opportunity to compare similar up and down regulated gene expression profiles, is needed. The main purpose of our work is to sort the extreme high and low expressed genes and extract, compare and cluster them. Expression profiles of 598 samples of invasive breast cancer and 48 samples of normal tissue have been analysed to create a new algorithm called Extreme Gene Expression Family (EGEF). The EGEF algorithm sorted, grouped and compared the highest and the lowest expressed genes (n = 100). According to the hierarchical clustering result, dense and light memberships of gene families are detected. The resulting analysis allows us to predict which genes would show similar expression signatures in invasive breast cancer and to us to recognize specific biological activities and processes. EGEF algorithm can be used to detect expression signatures in other cancers and biological processes.

**Keywords:** TCGA; Invasive breast cancer; EGEF algorithm; Expression pattern

## Introduction

The influence of changes in gene expression in the development of cancer is still not well understood [1]. Analysis of gene expression data in cancer studies is an effective way that leads to the discovery of global cancer profiling, tumor classification, tumor specific molecular marker identification and pathway exploration [2]. Gene expression profiling with next generation sequencing techniques has arisen as a powerful approach to study the cancer transcriptome [3]. This approach is valuable for the identification of novel biological mechanisms that are aberrant in cancer cells; moreover, this approach clarifies our understanding of known pathways in proteomics and the metabolome [4-6]. Next-generation sequencing has a big impact on cancer genomics in re-sequencing, analyzing, and comparing the tumor matches and normal genomes of a single patient [7]. The techniques have supplied large amounts of data about DNA sequencing, especially for cancer studies. Gene expression profiling with generation sequencing techniques has arisen as a powerful approach to study the cancer transcriptome. cDNA and oligonucleotide microarray technology have increased the rate of discovery of genetic interaction by simultaneously observing thousands of genes in a single experiment [4,8]. Gene expression approach is valuable to identify new mechanism in the regulation, expression and production of proteins and clarify our understanding of known pathways in the proteomics and the metabolome [5,9,10]. The cancer genomics area has been influenced profoundly by the application of next-generation sequencing technology, which has enormously speed up the pace of discovery while impressively decreasing the cost of data production [7]. Korbel first indicated that paired-end read from next generation sequencing platforms can be arranged to the genome and analyzed to determine Putative Structural Variation [11]. However, we need to digest this immense amount of gene expression data to turn into a sensible result about the genomics of cancer. There are several tools developed for this purpose. Oncomine is one of the most actively and statistically used cancer gene expression web tools. COPA [12,13] and GTI methods [14] are other methodss to be used statistically for cancer gene expression.

Here we introduce an algorithm, Extreme Gene Expression Family (EGEF), developed using the expression profiles of patients with invasive breast cancer in order to identify signatures that are characteristic for this cancer type. The main purpose of the algorithm is to find the highest and the lowest expressed genes and the correlation among them, specifically the genes which are coexpressed similarly in invasive breast cancer cells. The coexpression signatures of genes may elucidate novel mechanisms for the underlying biological processes in invasive breast cancer. The algorithm also allows us to detect the tumorigenensis involved genes and their sparse membership within the cancer.

## Material and Methods

### Subjects

Patients with invasive breast cancer (n = 598) and normal breast tissues (n = 48) were analyzed using microarray, UNC AgilentG4502A_0 [15]. Data on gene expression was downloaded from the TCGA data portal [16-18]. Each patient has the expression for 17814 genes.

### Data processing

Expression data from17814 genes for 598 invasive breast cancer samples and 48 normal breast tissues was downloaded from TCGA and applied to EGEF algorithm. Figure 1 shows the main steps in the work flow [19-22].

***Corresponding author:** Senol Dogan, Genetics and Bioengineering Department, International Burch University, Sarajevo, Bosnia and Herzegovina, Tel: +387 33 944 400; E-mail: senol1dogan3@gmail.com

**Received** February 01, 2016; **Accepted** February 27, 2016; **Published** March 07, 2016

**Citation:** Dogan S, Kurtovic-Kozaric A, Karlı G (2016) The Detection of Extremely High and Low Expressed Genes by EGEF Algorithm in Invasive Breast Cancer. J Biom Biostat 7: 286. doi:10.4172/2155-6180.1000286

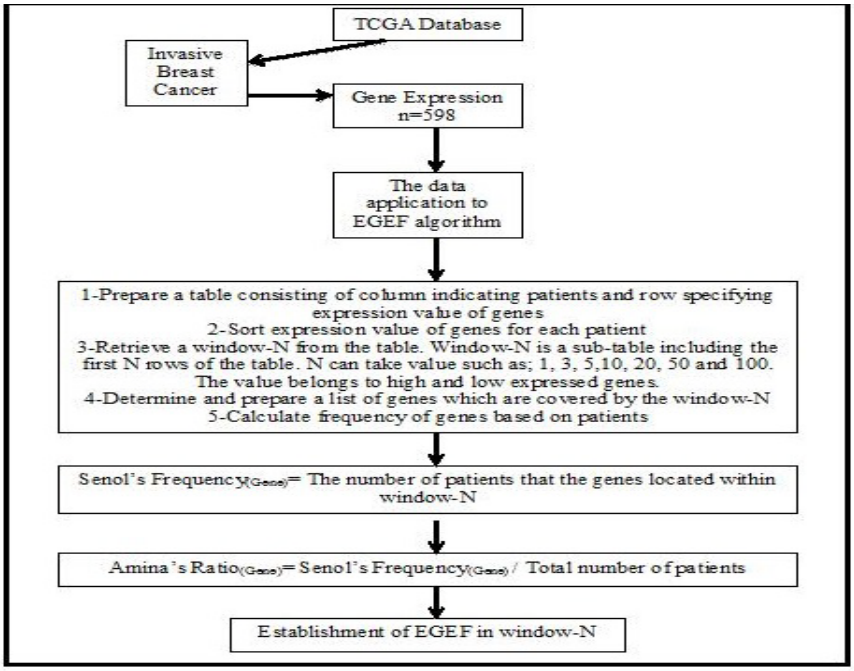


**Figure 1:** EGEF Algorithm Process. The data are downloaded and applied to the EGEF algorithm. The algorithm ends with each gene frequency and ratio. According to the frequency and ratio, Extreme Gene Expression Families have been produced.

## EGEF algorithm

The discovery of the highest (HE) and the lowest (LE) gene expression signatures is done by EGEF algorithm (Supplementary 1 and 2). The EGEF script has been created in R statistical program for fast and reliable data mining. EGEF sorts 598 samples with 17814 gene expression profiles and distinguishes extreme genes based on the expression level. We searched for top and bottom in set of 3, 5, 10, 20, 50 and 100 Hierarchical clustering, heat mapping, gene expression profiles and biological functions of genes are done by clustering and correlation programs HCE 3.0 and MSigDB respectively [9, 10].

## EGEF Algorithm steps (process)

1-Prepare a table consisting of column indicating patients and row specifying expression value of genes.

2-Sort expression value of genes for each patient ascending and descending order.

3-Retrieve a window-N from the table. Window-N is a sub-table including the first N rows of the table. N can take value such as 1, 3, 5,10, 20, 50 and 100.

4-Determine and prepare a list of genes which are covered by the window-N.

5-Calculate frequency of genes based on patients.

Senol's $Frequency_{(Gene)}$ = The number of patients that the genes located within window-N or first N

6-Calculate ratio of the genes. Higher ratio of the genes indicates the genes' activity in all patients.

Amina's $Ratio_{(Gene)}$ = Senol's $Frequency_{(Gene)}$/Total number of patients

7-Establish extremely expressed gene family which consist of the first n genes from window-N depending on their Amina's ratio.

## Implementation of EGEF algorithm

1-Prepare a table consisting of column indicating patients and row specifying expression value of genes. The algorithm starts with preparation of a table which consists of columns indicates patients and row specifying expression value of genes. To make it clear, four randomly genes were selected and prepared a table with real data (Table 1).

2-Sort expression value of genes for each patient ascending and descending order. The second step of the algorithm is to sort the value of genes for each patient. The sorting has been done in two ways; ascending and descending order (Table 2). The sorting will produce the highest and the lowest expressed genes [26-29].

3-Retrieve a window-N from the table. Window-N is a sub-table including the first N rows of the table. N can take value such as; 1, 3, 5,10, 20, 50 and 100. After the sorting of the data, a window-N has been retrieved from the table. The window-N is a sub-table which includes the first N rows of the table (Table 3). N can take value such as; 1, 3, 5,10, 20, 50 and 100. The window-N, N changeable, has been used to codify the extreme gene families.

4-Determine and prepare a list of genes which are covered by the window-N. The fourth step is to determine and prepare a list of genes which are covered by the window-N. The step is explained by two examples. The first example N = 3 and Patient = 598 and the second is N = 5 and Patient = 598; both are given in Table 4.

5-Calculate frequency of genes based on patients.

After all steps have been completed the algorithm has shown the

| | **Patient 1** | **Patient 2** | **Patient 3** | **…** | **Patient (n)=598** |
|---|---|---|---|---|---|
| Gene 1 | Expres.Value $_{(1.1)}$ | Expres.Value $_{(1.2)}$ | Expres.Value $_{(1.3)}$ | … | Expres.Value $_{(1.n)}$ |
| Gene 2 | Expres.Value $_{(2.1)}$ | Expres.Value $_{(2.2)}$ | Expres.Value $_{(2.3)}$ | … | Expres.Value $_{(2.n)}$ |
| Gene 3 | Expres.Value $_{(3.1)}$ | Expres.Value $_{(3.2)}$ | Expres.Value $_{(3.3)}$ | … | Expres.Value $_{(3.n)}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Gene(m) | Expres.Value $_{(m.1)}$ | Expres.Value $_{(m.2)}$ | Expres.Value $_{(m.3)}$ | | Expres.Value $_{(m.n)}$ |
| EXAMPLE | | | | | |
| ELMO2 | Patient 1 | Patient 2 | Patient 3 | ... | Patient (n)=598 |
| CREB3L1 | 1.55 | 2.89 | 3.46 | ... | 3.95 |
| RPS11 | 1.06 | 3.17 | 0.61 | ... | 5.39 |
| PNMA1 | 5.04 | 4.51 | 5.77 | ... | 3.89 |
| | ⋮ | ⋮ | ⋮ | ⋯ | ⋮ |
| Gene 17814 | 3.08 | 3.47 | 3.06 | | 2.60 |

EGEF algorithm works based on preparing tables. According to the genes n number patients expression value are written and prepared for the statistical analyzes. 17814 invasive breast cancer genes expression value from 598 patients are tabled in that order.

**Table 1:** Preparıng a table and example of that.

| | **Name of Gene** | **Patient 1** |
|---|---|---|
| Gene 3 | RPS11 | 5.04 |
| Gene 17814 | PNMA1 | 3.08 |
| Gene 1 | ELMO2 | 1.55 |
| Gene 2 | CREB3L1 | 1.06 |
| **Gene #** | **Name of Gene** | **Patient 1** |
| Gene 2 | CREB3L1 | 1.06 |
| Gene 1 | ELMO2 | 1.55 |
| Gene 17814 | PNMA1 | 3.08 |
| Gene 3 | RPS11 | 5.04 |

The algorithm is used for preparing tables depending on ascending and descending order to find top and bottom expressed genes respectively. The table shows how the genes are listed according to their expression value.

**Table 2:** Ascendıng and descendıng order of the genes.

| | **P 1** | **P 2** | **P 3** | **P 4** | **P 5** | **P 6** | **P 7** | **P 8** | **P 9** | **P 10** | **P…** | **P 598** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9.05 | 8.46 | 9.46 | 7.81 | 6.03 | 8.81 | 6.35 | 7.1 | 9.02 | 7.24 | … | 8.32 |
| 2 | 6.77 | 5.91 | 6.18 | 7.01 | 5.53 | 6 | 5.18 | 6.4 | 7.08 | 6.95 | … | 8.16 |
| 3 | 6.77 | 5.78 | 6.07 | 5.58 | 5.26 | 5.21 | 5.03 | 6.07 | 6.83 | 6.7 | … | 7.31 |
| 4 | 5.71 | 5.72 | 5.82 | 5.39 | 3.96 | 5.19 | 4.83 | 5.35 | 6.39 | 6.65 | … | 7.29 |
| 5 | 5.27 | 5.5 | 5.77 | 5.3 | 3.94 | 4.91 | 4.74 | 5.26 | 5.52 | 6.2 | … | 6.64 |
| 6 | 5.25 | 5.17 | 5.32 | 5.3 | 3.87 | 4.81 | 4.62 | 5.1 | 5.43 | 6.08 | … | 6.46 |
| 7 | 5.23 | 5.02 | 5.09 | 5 | 3.15 | 4.78 | 4.41 | 4.79 | 5.32 | 5.64 | … | 6.09 |
| 8 | 5.04 | 4.79 | 5.01 | 4.84 | 3.09 | 4.7 | 4.33 | 4.75 | 5.1 | 5.19 | … | 5.97 |
| 9 | 4.89 | 4.77 | 4.8 | 4.55 | 3.07 | 4.31 | 3.96 | 4.7 | 5.02 | 4.66 | … | 5.82 |
| 10 | 4.88 | 4.54 | 4.06 | 4.48 | 2.96 | 3.8 | 3.58 | 4.48 | 4.65 | 4.64 | … | 5.69 |
| 11 | 4.36 | 4.53 | 3.88 | 4.23 | 2.54 | 3.74 | 3.4 | 4.27 | 4.55 | 4.58 | … | 5.57 |
| 12 | 4.24 | 4.51 | 3.46 | 4.1 | 2.45 | 3.6 | 2.7 | 4.24 | 4.52 | 4.55 | … | 5.52 |
| 13 | 3.33 | 4.44 | 3.06 | 3.95 | 2.11 | 3.3 | 2.59 | 4.15 | 4.39 | 4.17 | … | 5.47 |
| 14 | 3.19 | 3.65 | 2.87 | 3.89 | 2.02 | 3.16 | 2.51 | 3.66 | 4.13 | 4 | … | 4.95 |
| 15 | 3.08 | 3.63 | 2.75 | 3.78 | 1.39 | 3.13 | 1.96 | 3.41 | 3.89 | 3.56 | … | 4.45 |
| 16 | 2.46 | 3.47 | 2.24 | 3.23 | 1.39 | 2.84 | 1.65 | 3.21 | 3.74 | 2.93 | … | 4.19 |
| 17 | 1.8 | 3.44 | 2.2 | 3.06 | 0.25 | 2.16 | 1.07 | 1.88 | 3.58 | 2.69 | … | 2.42 |
| 18 | 1.69 | 3.17 | 1.41 | 2.6 | 0.14 | 1.59 | 0.45 | 0.56 | 3.5 | 2.09 | … | 1.83 |
| 19 | 1.55 | 3 | 1.22 | 2.28 | -0.44 | 1.38 | -0.86 | 0.19 | 3.24 | -0.65 | … | 1.71 |
| 20 | 1.06 | 2.89 | 0.61 | 1.44 | -5.14 | 0.85 | -4.32 | -3.41 | 2.32 | -4.06 | … | -1.21 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 17814 | … | … | … | … | … | … | … | … | … | … | … | … |

The table shows Window-20. The expression value is sorted by descending order. P: Patient

**Table 3:** Wındow-n.

frequency of each gene based on patients. The frequency is the number of patients that the genes located within window-N. It is shown below and exampled.

Senol's Frequency$_{(Gene)w\text{-}N}$ = The number of patients that the genes located within window-N

For example, the number of patients that FCGR3A is located within window-50. So,

Senol's Frequency$_{(FCGR3A)w\text{-}50}$ = 589

6-Calculate the ratio of the genes. The frequency is used to find

the ratio of the genes if Senol's frequency is divided by total number of patients. As a result of that, special ratio, Amina's Ratio, has been found and show below with an example. The higher ratio of the genes indicates the genes intensive activity in all patients [29-31].

Amina's Ratio$_{(Gene)w\text{-}N}$ = Senol's Frequency$_{(Gene)w\text{-}N}$/Total number of patients

Example: Senol's frequency of FCGR3A is 589 and total number of patients = 598

Amina's Ratio $_{(FCGR3A)w\text{-}50}$ = 589/598 = 0.98

7- Establish extremely expressed gene family consisting of the first n genes from window-N depending on their Amina's ratio.

## Similar algorithms

According to their target and performance, similar algorithms which search for the gene expression patterns and clustering methods have been cited (Table 5).

| | Gene | Expression Value |
|---|---|---|
| Patient 1 | ASPN | 9.05 |
| Patient 1 | CRH | 9.04 |
| Patient 1 | MUM1L1 | 7.81 |
| Patient 2 | ASPN | 8.46 |
| Patient 2 | SCGB2A2 | 7.86 |
| Patient 2 | CPA3 | 6.60 |
| … | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| | | |
| | **Gene** | **Expression Value** |
| Patient 1 | ASPN | 9.05 |
| Patient 1 | CRH | 9.04 |
| Patient 1 | MUM1L1 | 7.81 |
| Patient 1 | RGS1 | 7.48 |
| Patient 1 | COL11A1 | 7.45 |
| Patient 2 | ASPN | 8.46 |
| Patient 2 | SCGB2A2 | 7.86 |
| Patient 2 | CPA3 | 6.60 |
| Patient 2 | EHF | 6.60 |
| Patient 2 | MMP7 | 6.55 |
| … | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| Patient 598 | … | … |
| The genes are applied to the algorithm by using Window-N. The patients top 3 and top 5 genes are listed respectively. Changing the N number in Window different or common top and bottom expresssed genes are detected. | | |

**Table 4:** Genes ıncluded by wındow-3 and wındow-5.

| Algorithm | Types | Reference |
|---|---|---|
| Agglomerative HCT | Investigate any correlation among discriminator genes in hereditary breast cancer | [11] |
| E-cast | Uses a dynamic threshold | [12] |
| Non-HCT (CAST) | Clustering gene expression patterns. | [13] |
| Bayesian Biclustering | Searches for local patterns of gene expression | [14] |
| FABIA | Accounts for linear dependencies between gene expression and conditions | [15] |
| QUBIC | Combination of qualitative measures of gene expression data and a combinatorial optimization technique | [16] |
| CPB | A comparative analysis of biclustering algorithms for gene expression data | [17] |
| Combinatorial Clustering | Three classification techniques comparison, k-NN,SVM and AdaBoost | [18] |
| Worst-Case | Worst-Case Analysis of Selective Sampling for Linear Classification | [19] |
| COALESCE | Co-regulated and sequence-level regulatory motifs | [20] |
| Cheng and Church | Biclustering of expression data | [21] |
| Plaid | A tool for exploratory analysis of multivariate data | [22] |
| BiMax | Sharing compatible expression patterns across subsets of samples | [23] |
| xMOTIFs | A conserved gene expression motif | [24] |
| OPSM | Capturing the general tendency of gene expressions across a subset of conditions | [25] |
| Spectral MEQPSO | Global convergence towards an optimal solution | [26] |
| ISA | Overlapping transcription modules | [27] |

**Table 5:** Related clustering algorıthms.

## Results

### EGEF analysis for top and bottom expressed genes

A new value termed Senol's frequency, was created to refer to the number of window-N. According to the Senol's frequency, the highest and the lowest expressed genes of invasive breast cancer have been found. The frequency is calculated for the highest and lowest expressed genes in window-50 and are given in Table 5 and Table 6, respectively. In addition, the 100 genes in window-100 are available in Supplementary 1. For example, Senol` frequency$_{(FCGR3A)w\text{-}50}$ = 589 means the gene expression is located at top of window-50 (589 out of 598 patients have the expression of the specified gene within the top 50 highest expressed genes). Amina's ratio provides the information regarding the particular gene expression activity in the breast cancer. For example, if Senol's frequency for FCGR3A is 589 and the total number of patients is 598, then Amina's Ratio is (FCGR3A)$_{w\text{-}50}$ is 0.98. After the application of the EGEF algorithm, the extreme HE and LE genes are grouped as top and bottom window-N, N = 3, 5, 10, 25, 50, 100 members (Tables 6 and 7).

### The algorithm selected genes compared with control data

According to the frequency, the patients gene expression have been compared to the control data in order to observe the difference of expression activity. Only 25 of 100 extreme the highest and the lowest genes are presented in this paper (Figure 2). For example, ASPN expression average is 7.21 in the cancer cell, but it is too low in the control expression, -0.03 (Figure 2). The same result has been seen in all of the highly expressed gene levels. The algorithm has selected genes that show that there is a large difference in expression between patient and control samples in the high expressed group. However, the comparison of low expressed genes in patient and control samples shows almost no difference. For example, AHSG has very low gene expression frequency (Senol frequency = 593). The gene expression level for AHSG gene drops from -6.29 to -6.59 between cancer and control cell (Figure 2).

### The selected gene's involvement of tumorigenesis and cellular activity

100 extreme high and the low expressed genes are categorized depending on their biological features (Supplementary 4) so that we can predict what kind of mechanisms are potentially activated in the breast cancer cell. Based on the function of the specific gene, we have searched for the potential of the genes' involvement in tumorigenesis (Table 8). We have observed the following correlation regarding the size of the window-N and the involvement of a specific gene in tumorigenesis—the smaller the window-N, the stronger the relation with tumorigenesis. Similarly, the larger the window-N, the weaker the relation in the highly expressed genes. If the value of N takes 3,5 and 10, tumorigenesis involvement is 100%. If N takes 20 or 50, the involvement is 84% and 77%, respectively (Table 8). The low expressed genes have shown different tumorigenesis relation percentage. The highest tumorigenesis involvement detected in window-N = 3 is 100%, but after that the relation to tumorigenesis decreases. N takes the value of 5, 10, 20, 50 and percentage is 60%, 60%, 65% and 62% respectively (Table 8).

### Clustering of the EGEF selection genes

The 50 genes which are extremely high and low expressed are clustered respectively, and shown in a heat map (Figures 3 and 4). The Pearson correlation is used to calculate the gene expression correlation.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | **FCGR3A** | 589 |
| | | | | | | | | **CD163** | 586 |
| | | | | | | **FCGR3A** | 552 | **RGS1** | 547 |
| | | | | | | **CD163** | 543 | **ASPN** | 541 |
| | | | | | | **ASPN** | 495 | **SCGB2A2** | 468 |
| | | | | | | **RGS1** | 470 | **MMP7** | 456 |
| | | | | | | **SCGB2A2** | 441 | **COL3A1** | 451 |
| | | | | **ASPN** | 435 | **EHF** | 337 | **PIP** | 443 |
| | | | | **CD163** | 405 | **MMP7** | 327 | **GGTA1** | 435 |
| | | | | **SCGB2A2** | 404 | **CXCL9** | 298 | **EHF** | 424 |
| | | **SCGB2A2** | 374 | **FCGR3A** | 382 | **GRP** | 276 | **CXCL9** | 424 |
| SCGB2A2 | 337 | **ASPN** | 362 | **RGS1** | 371 | **GGTA1** | 272 | CLEC14A | 394 |
| ASPN | 312 | **RGS1** | 271 | **EHF** | 229 | COL11A1 | 263 | CPA3 | 385 |
| RGS1 | 207 | **CD163** | 174 | **MMP7** | 182 | **PIP** | 263 | COL10A1 | 384 |
| | | **EHF** | 137 | **GRP** | 173 | **SCGB2A1** | 250 | FGL2 | 380 |
| | | | | **CXCL9** | 168 | CPA3 | 235 | SFRP4 | 374 |
| | | | | **SCGB2A1** | 168 | **SCGB1D2** | 219 | **GRP** | 365 |
| | | | | | | CYP4Z1 | 212 | COL11A1 | 354 |
| | | | | | | ADAMDEC1 | 189 | **SCGB1D2** | 340 |
| | | | | | | COL10A1 | 187 | **FNDC1** | 338 |
| | | | | | | **FABP4** | 184 | **SCGB2A1** | 331 |
| | | | | | | **DACH1** | 181 | CILP | 327 |
| | | | | | | | | **FABP4** | 296 |
| | | | | | | | | **CD93** | 289 |
| | | | | | | | | **DACH1** | 286 |
| | | | | | | | | **WNT2** | 166 |
| | | | | | | | | **OLFM4** | 165 |

The genes are selected according to the Window-N, 3, 5, 10, 25 and 50. The last column belongs to Window-50 and the lats two rows represents 49$^{th}$ and 50$^{th}$ top genes and their frequency. Bold genes stand for tumorigenesis involvement.

**Table 6:** The high expressed genes and their senol's frequency.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | **TYR** | 593 |
| | | | | | | | | AHSG | 593 |
| | | | | | | **TYR** | 584 | **IGFBP1** | 589 |
| | | | | | | **IGFBP1** | 577 | RPS4Y1 | 589 |
| | | | | | | RPS4Y1 | 575 | **FABP1** | 579 |
| | | | | | | AHSG | 560 | SERPINA7 | 572 |
| | | | | | | NTS | 528 | NTS | 556 |
| | | | | **IGFBP1** | 572 | **FGL1** | 466 | APOH | 554 |
| | | | | RPS4Y1 | 492 | **FABP1** | 465 | **HEMGN** | 549 |
| | | | | **TYR** | 452 | FGB | 441 | RPS4Y2 | 543 |
| | | **IGFBP1** | 572 | NTS | 424 | **MAGEC2** | 427 | **FGL1** | 527 |
| IGFBP1 | 560 | RPS4Y1 | 492 | **FABP1** | 402 | UTS2 | 383 | **SPANXD** | 525 |
| FABP1 | 195 | **TYR** | 452 | AHSG | 367 | **SPANXD** | 348 | TTR | 523 |
| TFPI2 | 161 | NTS | 424 | **FGL1** | 360 | **DSCR8** | 315 | **HMGA2** | 521 |
| | | **FABP1** | 402 | FGB | 349 | TTR | 315 | DPPA4 | 508 |
| | | | | **MAGEC2** | 325 | **TFPI2** | 287 | **MAGEC2** | 498 |
| | | | | **TFPI2** | 233 | **PTX3** | 286 | **FGB** | 486 |
| | | | | | | SERPINA7 | 267 | **ITIH2** | 485 |
| | | | | | | **MAGEC1** | 257 | UTS2 | 483 |
| | | | | | | **HMGA2** | 255 | **MAGEC1** | 450 |
| | | | | | | **APOH** | 249 | **DSCR8** | 433 |
| | | | | | | **DKK1** | 229 | HBG1 | 423 |
| | | | | | | | | CPB2 | 413 |
| | | | | | | | | NR0B1 | 400 |
| | | | | | | | | C16orf73 | 383 |
| | | | | | | | | MTTP | 223 |
| | | | | | | | | DDX3Y | 222 |

The genes are selected according to the Window-N, 3, 5, 10, 25 and 50. The last column belongs to Window-50 and the last two rows of that represents 49th and 50th bottom genes and their frequency. Bold genes stand for tumorigenesis involvement.

**Table 7:** The low expressed genes and their senol's frequency.

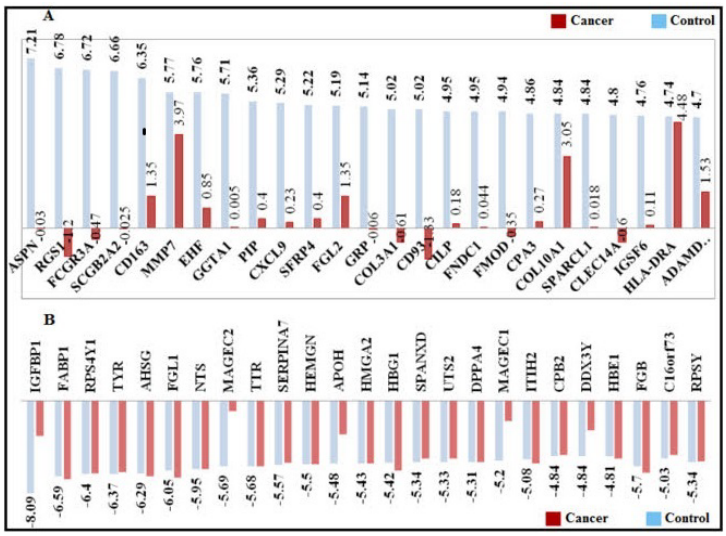


**Figure 2:** Comparison of Cancer and Control Gene Expression. Table A is the comparison of 25 the highest cancer and control gene expression and the average of them is 5A2 and 0.48 respectively. Table B compares 25 the lowest cancer and control gene expression and the average is almost same -4.98. While the high expressed genes average is too high the low expressed genes average is almost same in the breast cancer.

| **Window-3** | **Window-5** | **Window-10** | **Window-20** | **Window-50** |
|---|---|---|---|---|
| **100%** | **100%** | **100%** | **84%** | **77%** |
| | | | | B |
| Window-3 | Window-5 | Window-10 | Window-20 | Window-50 |
| **100%** | **60%** | **60%** | **65%** | **62%** |

According to Senol's frequency (A) shows percentage of the top expressed genes which are related with tumorigenesis. (B) shows percentage of the bottom expressed genes which are related with tumorigenesis.

**Table 8:** Tumorigenesis involvement of window-NA.

**Figure 3:** EGEF high expressed genes. The heat map is generated by HCE 3.0 and clustered according to Pearson correlation. The figure shows the relation of the 50 extreme gene expressions among each other. Among the high expressed genes there is clear diversity. Although that the number genes which are expressed together is very high.

**Figure 4:** EGEF low expressed genes. The figure is generated by HCE 10 and clustered according to Pearson correlation. The heat map shows 50 low expressed genes clustering and relations. AHSG, TYR, FABPI, RPS4YI, NTS show strong correlation as in the low expressed genes in the cancer. The figure quite clear shows the strong and weak correlated genes.

The heat maps show us the clustered and correlation among 50 extreme high and low genes [31-34].

## Discussion

The changes in global gene expression lead us to understand better of the biological activities which drive to carcinogenesis. The EGEF algorithm sorted, grouped and compared the highest and the lowest expressed genes (n = 100). The resulting analysis allows us to predict which genes would show similar expression signatures in invasive breast cancer, allowing us to recognize specific biological activities and processes. EGEF algorithm can be used to detect expression signatures in other cancers and biological processes. In the future, the results of the EGEF algorithm can be correlated with clinical parameters in order to find potential new targets for drug treatment targets. Most of the algorithms focus on finding the outlier expressed genes, oncogenes or tumor supressor genes, but the EGEF algorithm points out tumorigenesis related genes and their partner genes that help a cell to convert to cancer cell Clinical and genomic work regarding cancer need a new perspective to look at the heterogenity of the cancer development and clinical treatment. The new algorithm takes a different approach than previous approaches which only target abnormally expressed genes. However, the main goal of EGEF algorithm is to find tumoigenesis related genes and their family members and their relation strength of the family. If we change our view of the problem, then we might be able to find new solutions or ways to target therapy.

### Author Disclosure Statement

The authors declare that no competing financial interests exist.

### Acknowledgement

We would like to thank Association for the Advancement of Science in Bosnia and Herzegovina and Bosnia Sema Education for their support.

### References

1. Nevins JR, Potti A (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. Nat Rev Genet 8: 601-609.
2. Fung BYM (2004) Meta-classification of multi-type cancer gene expression data. Proceeding of 4th Workshop on Data Mining in Bioinformatics p: 31-39.

3. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66-72.
4. Arakawa K, Kono N, Yamada Y, Mori H, Tomita M (2005) KEGG-based pathway visualization tool for complex omics data. In Silico Biol 5: 419-423.
5. Bono H, Okazaki Y (2005) The study of metabolic pathways in tumors based on the transcriptome. Semin Cancer Biol 15: 290-299.
6. Howbrook DN, van der Valk AM, O'Shaughnessy MC, Sarker DK, Baker SC, et al. (2003) Developments in microarray technologies. Drug Discov Today 8: 642-651.
7. Clarke JD, Zhu T (2006) Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives. The Plant Journal 45: 630-650.
8. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, et al. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. Lancet 359: 1301-1307.
9. Howbrook DN, van der Valk AM, O'Shaughnessy MC, Sarker DK, Baker SC, et al. (2003) Developments in microarray technologies. Drug Discov Today 8: 642-651.
10. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318: 420-426.
11. Furusato B, Gao CL, Ravindranath L, Chen Y, Cullen J, et al. (2008) Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer. Mod Pathol 21: 67-75.
12. Tibshirani R, Hastie T (2007) Outlier sums for differential gene expression analysis. Biostatistics 8: 2-8.
13. Mpindi JP, Sara H, Haapa-Paananen S, Kilpinen S, Pisto T, et al. (2011) GTI: a novel algorithm for identifying outlier gene expression profiles from integrated microarray datasets. PLoS One 6: e17259.
14. http://www.chem.agilent.com/cag/bsp/gene_lists.asp
15. (2015) The Cancer genome atlas.
16. Seo J, Bakay M, Chen YW, Hilmer S, Shneiderman B, et al. (2004) Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. Bioinformatics 20: 2534-2544.
17. http://software.broadinstitute.org/gsea/msigdb/collections.jsp
18. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. (2001) Gene-expression profiles in hereditary breast cancer. N Engl J Med 344: 539-48.
19. Abdelghani B, David P, Yidong C, Abdel GE (2002) E-CAST: A data mining algorithm for gene expression data. BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference) p 49-54.
20. Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering gene expression patterns. J Comput Biol 6: 281-297.
21. Gu J1, Liu JS (2008) Bayesian biclustering of gene expression data. BMC Genomics 9 Suppl 1: S4.
22. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, et al. (2010) FABIA: factor analysis for bicluster acquisition. Bioinformatics 26: 1520-1527.
23. Li G, Ma Q, Tang H, Paterson AH, Xu Y (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Res 37: e101.
24. Eren K, Deveci M, Kuçuktunç O, Çatalyurek UV (2013) A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform 14: 279-292.
25. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406: 536-540.
26. Nicolo CB, Claudio G, Luca Z (2006) Worst-case analysis of selective sampling for linear classification. Journal of Machine Learning Research 7: 1205-1230.
27. Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, et al. (2009) Detailing regulatory networks through large scale data integration. Bioinformatics 25: 3267-3274.
28. Cheng Y, Church GM (2000) Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 8: 93-103.
29. Laura L, Art O (2002) Plaid models for gene expression data. Statistica Sinica 12: 61-86.
30. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22: 1122-1129.
31. Murali TM, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. Pac Symp Biocomput.
32. Byron JG, Obi LG, Martin E, Steven JJ (2006) Discovering Significant OPSM Subspace Clusters in Massive Gene Expression Data. KDD.
33. Vijayalakshmi S, Rajalakshmi MJ, Jayanavithraa C, Ramya L (2013) Gene expression data analysis using automatic spectral MEQPSO clustering algorithm. International Journal of Advanced Research in Computer and Communication Engineering.
34. Bergmann S, Ihmels J, Barkai N (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys 67: 031902.