International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:3, No:10, 2009

Identification of Complex Sense-antisense Gene's Module on 17q11.2 Associated with Breast Cancer Aggressiveness and Patient's Survival

O. Grinchuk, E. Motakis and V. Kuznetsov

***Abstract*—**Sense-antisense gene pair (SAGP) is a pair of two oppositely transcribed genes sharing a common region on a chromosome. In the mammalian genomes, SAGPs can be organized in more complex sense-antisense gene architectures (CSAGA) in which at least one gene could share loci with two or more antisense partners. Many dozens of CSAGAs can be found in the human genome. However, CSAGAs have not been systematically identified and characterized in context of their role in human diseases including cancers. In this work we characterize the structural-functional properties of a cluster of 5 genes –TMEM97, IFT20, TNFAIP1, POLDIP2 and TMEM199, termed TNFAIP1 / POLDIP2 module. This cluster is organized as CSAGA in cytoband 17q11.2. Affymetrix U133A&B expression data of two large cohorts (410 patients, in total) of breast cancer patients and patient survival data were used. For the both studied cohorts, we demonstrate (i) strong and reproducible transcriptional co-regulatory patterns of genes of TNFAIP1/POLDIP2 module in breast cancer cell subtypes and (ii) significant associations of TNFAIP1/POLDIP2 CSAGA with amplification of the CSAGA region in breast cancer, (ii) cancer aggressiveness (e.g. genetic grades) and (iv) disease free patient's survival. Moreover, gene pairs of this module demonstrate strong synergetic effect in the prognosis of time of breast cancer relapse. We suggest that TNFAIP1/POLDIP2 cluster can be considered as a novel type of structural-functional gene modules in the human genome.

***Keywords*—**Sense-antisense gene pair, complex genome architecture, TMEM97, IFT20, TNFAIP1, POLDIP2, TMEM199, 17q11.2, breast cancer, transcription regulation, survival analysis, prognosis.

I. Introduction

A cis-sense antisense gene pair (SAGP) is a pair of genes mapped to opposite strands on the same locus and therefore transcribed in opposite directions. Corresponding pairs of cis-antisense transcripts are mRNAs that are at least partially complementary to each other. Cis-antisense mRNAs that are naturally transcribed from SAGP are known as naturally occurred sense-antisense (SA) RNAs. Such SA transcripts (SAT) have been observed in prokaryotes, fungi, plants, and animals [1]-[4]. The overlapping of protein coding genes is quite a common feature of SAGP in prokaryotic genomes [5], [6]. However, up to 32% of yeast genes [1] and up to 25% of mammalian genes [3] have been estimated to reside in SATs.

Natural SAT have already been found to function at several levels of molecular eukaryotic gene regulation including alternative initiation, splicing, termination [7], translational regulation [8], RNA stability, trafficking, apoptosis [9],[2] genomic imprinting [10], antisense mediated silencing [11] and development including X-inactivation [12], eye development [13].

Case studies have showed that changes in SAGPs transcription could be implicated in pathological processes such as some neurology and cancer diseases [11], [14]. It was shown experimentally, that in leukemia cells, SA gene pair BAL1 and BBA is bi-directionally transcribed, concordantly expressed due to INF-gamma induction and can directly interact at the protein level [15]. In our work [16], we have reported 12 high-confidence SAT pairs which are concordantly regulated in human breast cancer tissues. Among these gene pairs two pairs (RAF1-MKRN2 and CKAP1-POL2I) are co-regulated constitutively in breast tumors of different genetic grades (G1,G1-like, G3-like, and G3), while the expression of CR590216-EAP30 SA gene pair is gradually changed with aggressiveness of cancer in genetic groups G1,G1-like, G3-like, and G3, respectively.

In the mammalian genomes, SAGPs can be organized in more complex sense-antisense gene architectures (CSAGA) in which at least one gene could share loci with two or more antisense partners [17], [18]. It has been shown in several case studies that SAGPs may be involved in cancer and some other diseases and could be associated with complex disease syndromes. Many dozens CSAGA can be found in the human genome [3], [18], [19]. However, CSAGAs have not been characterized in context of their role in human diseases including cancers.

A gene density on the human chromosome 17 is relatively higher than on the most other human chromosomes. There are many oncogenes on human Chr17. Nevertheless, localization of these genes is not uniform. For example, based on Cancer GeneticsWeb (www.cancer-genetics.org), the oncogenes TAF2N, NF1, THRA and ERBB2 are located in 17q11.1-q12. ERBB2 (*Her-2/neu*), a well-known oncogene, is located on 17q12. Many other genes located close to *ERBB2* on 17q12 could be over-expressed / amplified and are known or suspected to play a role in carcinogenesis, and, specifically, in breast carcinogenesis. Previous studies demonstrated that the negative effect on the prognosis of breast cancer attributed to

ERBB2 amplification could, in fact, be due to co-amplification of the region adjacent to ERBB2 [20]. The ERBB2 and its neighbor genes could be amplified and over-expressed in 25% of invasive breast carcinomas [21], [22]. In general, ERBB2 amplification and over-expression confers an unfavorable prognosis, although its significance is less than that of the traditional prognostic factors – stage and grade. Also, it seems that the prognosis and response to therapy varies considerably within the spectrum of *ERBB2*-amplified breast carcinomas, indicating that they are biologically heterogeneous [22].

CSAGAs and their association with human cancers in the regions outside of the ERBB2 amplicon core region in 17q12 [23] have not been studied.

In this work we suggested that a high diversity of breast cancer cell subtypes could be probably associated with the active chromatin regions located on 17q, different from the ERBB2 amplicon region. We also assume that novel CSAGAs could be found on 17q and these complex regions could play significant role in transcription control resulting in cancer phenotypes and patient's survival.

II. Materials and Methods

Patients tumor specimens, cell lines and microarray data.

Clinical characteristics of breast cancer patients and tumor samples of two independent cohorts (Uppsala and Stockholm) were published in [24]. Stockholm cohort comprised of K_s = 159 patients with breast cancer, operated in Karolinska Hospital from 1st of January 1994 to 31st of December 1996, identified in the Stockholm-Gotland breast Cancer registry [24]. Uppsala cohort involved K_u = 251 patients representing approximately 60% of all breast cancers resections in Uppsala County, Sweden, from 1 January 1987 to 31 December 1989. Information on patients' disease-free survival (DFS) times/events and the expression patterns of approximately 30,000 gene transcripts (representing N = 44,928 probe sets on Affymetrix U133A and U133B arrays) in primary breast tumors were obtained from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (Stockholm data set label is GSE4922; Uppsala dataset label is GSE1456). The microarray intensities were MAS5.0 calibrated and the probe set signal intensities were log-transformed and scaled by adjusting the mean signal to a target value of log500.

Correlation analysis

We were going to identify whether a significant cluster is formed among certain genes organized in a CSAA on chromosome position 17q11.2. Our first step to show the existence of such a cluster was to estimate Kendall correlations among these genes in the two large cohorts and subsequently test whether their respective matrix is significant at level α = 1%. Formally, the correlation matrix we derive is:

$$R = \begin{pmatrix} 1 & \cdots & r_{p1} \\ \vdots & \ddots & \vdots \\ r_{1p} & \cdots & 1 \end{pmatrix}$$

where r_{1p} denotes the Kendall correlation coefficient between affy probes 1 and p, estimated from the microarray expression data, and p is the total number of probes in the prospective cluster.

To test the significance of R matrix, we used a bootstrap version of Barlett's statistical test [25]. Bootstrap Bartlett test evaluates the significance of the hypothesis H_0: $R_{pxp} = I_{pxp}$, where R_{pxp} is the p x p correlation matrix and I_{pxp} is the corresponding p x p identity matrix. Under null hypothesis, there is no significant correlation among these probes, whereas rejection of H_0 at α = 1% is an indication of a cluster. For the p genes of the correlation matrix one needs to compute the statistic:

$$T = -[(N-1) - \frac{1}{6}(2p+5)]\log|R|,$$

where N is the sample size (number of patients in each cohort), p is the number of variables (probes) and $|R|$ is the determinant of the sample correlation matrix. This quantity is distributed approximately as χ^2 with ½ $p(p-1)$ degrees of freedom. To test the significance of the statistic, draw B = 5000 samples of k genes at random from the set of 44928 genes and estimate Bartlett's T-test, T^b, for each b = 1, 2, …, 5000 draws. The corresponding bootstrap p-value is estimated as: p^{boot} = number of times $\{T^b \geq T\}/B$. Similar bootstrap approaches have been discussed in [26].

Comparison of correlation matrices

We would like to show that the genes in R matrix are more strongly correlated than any other neighboring genes. In this way we show that they form a significant, tight cluster that cannot be re-produced in the neighborhood. For our analysis we use Box' M test [27], which evaluates the significance of the hypothesis H_0: $R_{pxp} = R^*_{qxq}$, where R_{pxp} is as before and R^*_{qxq} is a neighboring qxq correlation matrix, where $q \neq p$ or $q = p$. To compute the test, we calculate:

$$M = \frac{|S_p|^{p/2} \times |S_q|^{q/2}}{|S_{pool}|^{(p+q)/2}},$$

where $|S_p|$ is the determinant of the variance-covariance matrix of our prospective gene cluster (corresponding to R_{pxp} correlation matrix), $|S_q|$ is the determinant of the variance-covariance matrix of the neighbor group of genes (corresponding to R^*_{qxq} correlation matrix) and $|S_{pool}|$ is the pooled sample variance/covariance matrix estimated as:

$$S_{pool} = \frac{p \times S_p + q \times S_q}{p+q}$$

Box [27] gave χ^2 and F approximations for the distribution of M (exact test does not exist).

Survival Analysis Based on Genes and Gene Pairs Expression Patterns.

This analysis involves testing whether the prospective gene cluster contains *survival significant genes* and *survival synergistic gene pairs*. To this extent, we briefly describe the survival analysis approach we apply to our data. Detailed discussion can be found in [28].

We assume a microarray experiment with i = 1, 2, ..., p genes, whose log-transformed intensities are measured for k = 1, 2, ..., K patients. Associated with each patient are a continuous clinical outcome data (Disease Free Survival time t_k; defined as the time interval from surgery until the first

recurrence (local, regional, distant) or last date of follow up), and a nominal (yes/no) clinical event e_k (occurrence of tumor metastasis at time t_k). Each patient is assigned to the low- or high- risk group according to:

$$x^i{}_k = \begin{cases} 1 \text{ (high-risk)}, \text{if } \ y_{i,k} > c^i \\ 0 \ \text{ (low-risk)}, \text{if } \ y_{i,k} \le c^i \end{cases},$$

where c^i denotes the cutoff of the ith gene's intensity level. Motakis and coauthors [28] showed how to estimate this cutoff from the data using the 1 dimensional data-driven grouping algorithm (1D DDg). The clinical outcomes/events are subsequently fitted to the patients' groups by the Cox proportional hazard regression model [29]:

$$\log h^i{}_k(t_k \mid x^i_k, \beta_i) = \alpha_i(t_k) + \beta_i \cdot x^i_k \ ,$$

where $h^i{}_k$ is the hazard function and $\alpha_i(t_k) = \log h^i{}_0(t_k)$ represents the unspecified log-baseline hazard function; β is the $1 \times p$ regression parameters vector; and t_k is the patients survival time. To assess the ability of each gene to discriminate the patients into two distinct genetic classes, the Wald P value of the β_i coefficient of model [29] is estimated by using the univariate Cox partial likelihood function, estimated for each gene i as:

$$L(\beta_i) = \prod_{k=1}^{K} \left\{ \frac{\exp(\beta_i^T x^i_k)}{\sum_{j \in R(t_k)} \exp(\beta_i^T x^i_j)} \right\}^{e_k} ,$$

where $R(t_k) = \{j: t_j \ge t_k\}$ is the risk set at time t_k and e_k is the clinical event at time t_k. The actual fitting of Cox model is conducted by the *survival* package in R (http://cran.r-project.org/web/packages/survival/index.html). The genes with the smallest β_i Wald p-values are assumed to have better group discrimination ability and thus called *highly survival significant genes*. These genes are selected for further confirmatory analysis or for inclusion in a prospective gene signature set.

Similar approach is applied to identify synergistic survival significant gene pairs using the 2 dimensional data-driven grouping method of Motakis et al., 2007[28]. Briefly, for a given gene pair i, j with individual cutoffs (identified from the 1 dimensional data-driven grouping) c^i and c^j, $i \ne j$, we may classify the K patients by seven possible two-group designs. Figure 1 indicates the regions where patients' gene intensities

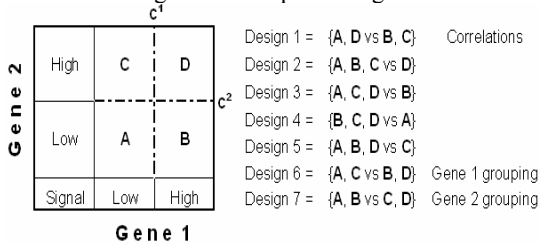

Fig. 1. Grouping of a synergetic gene pair (genes 1 and 2 with respective cutoffs c^1 and c^2) and all possible two-group designs (Designs 1-7).

$[y_{i,k}, y_{j,k}]$ are plotted; mind that "A", "B", "C" and "D" are defined by the conditions $y_{i,k} < c^i$ and $y_{j,k} < c^j$; $y_{i,k} \ge c^i$ and $y_{j,k} < c^j$; $y_{i,k} < c^i$ and $y_{j,k} \ge c^j$; $y_{i,k} \ge c^i$ and $y_{j,k} \ge c^j$.

For $i = 1$ and $j = 2$, group the patients by each of the seven designs of Figure 1 (using individual gene cutoffs), fit Cox model for each design and estimate the seven Wald P values for β_i. Provided that the respective groups sample sizes are sufficiently large and the assumptions of model (25) are satisfied, the best grouping scheme among the five "synergetic" (1 – 5) and the two "independent" (6 – 7) designs is the one with the smallest β_i P value. Iterate 1 for all i and j combinations of the k genes ($i = 1, \ldots, p - 1, j = i + 1, \ldots, p$).

Correlation and survival analysis have been conducted in R (http://cran.r-project.org/) using software developed by our group. All our programs are available upon request.

III. RESULTS

Identification of co-expressed TNFAIP1-POLDIP2 SA gene pair

Using high-confidence Affymetrix Chip U133 A&B probesets presented in APMA database [30] (http://apma.bii.a-star.edu.sg/) we selected 156 SA transcripts pairs located on chromosome 17 with reliable RefSeq support (NM ID) for each member of a pair. We focused on chromosome 17 based on 2 known facts: 1) many regions of chromosome 17 are actively involved in recurrent amplifications during breast cancer development (including ERBB2 amplicon), [31], and 2) gene density on the human chromosome 17 is relatively higher than on the other human chromosomes, except chromosome 19 [32].

Next, utilizing expression data from Uppsala and Stockholm cohorts we calculated Kendal correlations for each pair and selected high-confidence top level correlated pairs of probesets representing SA gene pairs. Among positive high-correlated SA pairs, two genes TNFAIP1 and POLDIP2 are selected and studied in this work. POLDIP2 (NM_015584) gene encodes a protein which interacts with the DNA polymerase delta p50 subunit and with proliferating cell nuclear antigen (PCNA) [42]. Some transcripts of this gene overlap in a tail-to-tail orientation with the gene for tumor necrosis factor (TNF) alpha-induced protein 1, TNFAIP1 (NM_021137). The genes of this pair form a convergent (tail-to-tail) gene orientation topology, sharing 376-nt region of their 3'UTRs and are located on human chromosome cytoband 17q11.2. It has been reported that this gene can be induced by TNF-alpha [52]. Moreover, TNFAIP1 protein can directly interact with PCNA protein. In the rat, TNFAIP1gene could stimulate polymerase delta activity in vitro in PCNA-dependent way [47]. Thus, transcription of POLDIP2 and TNFAIP1 genes can be under common control and products of these genes can be involved in the same pathways.

Identification of TNFAIP1/POLDIP2 SFGM

We identified a convergent TNFAIP1/POLDIP2 SAGP located on 17q11.2. On one hand, this SAGP demonstrated reproducible and significant co-expression pattern in 2 independent cohorts (Uppsala and Stockholm Cohorts (see Methods)) of breast cancer patients; on the other hand, when

these genes were analyzed as a pair, they turned out to be survival significant in both cohorts (see materials and methods section).

Affymetrix U133A&B probes 214283_at (gene TMEM97), 229182_at as well as 233531_at and 234060_at (gene SLC46A1) were excluded from our analysis due to their unclear support by the well annotated and reliable RefSeq gene model. Genes (and, correspondingly, Affymetrix probes) in the matrix were placed in the same order as they are located on 17q11.2 in the genome.

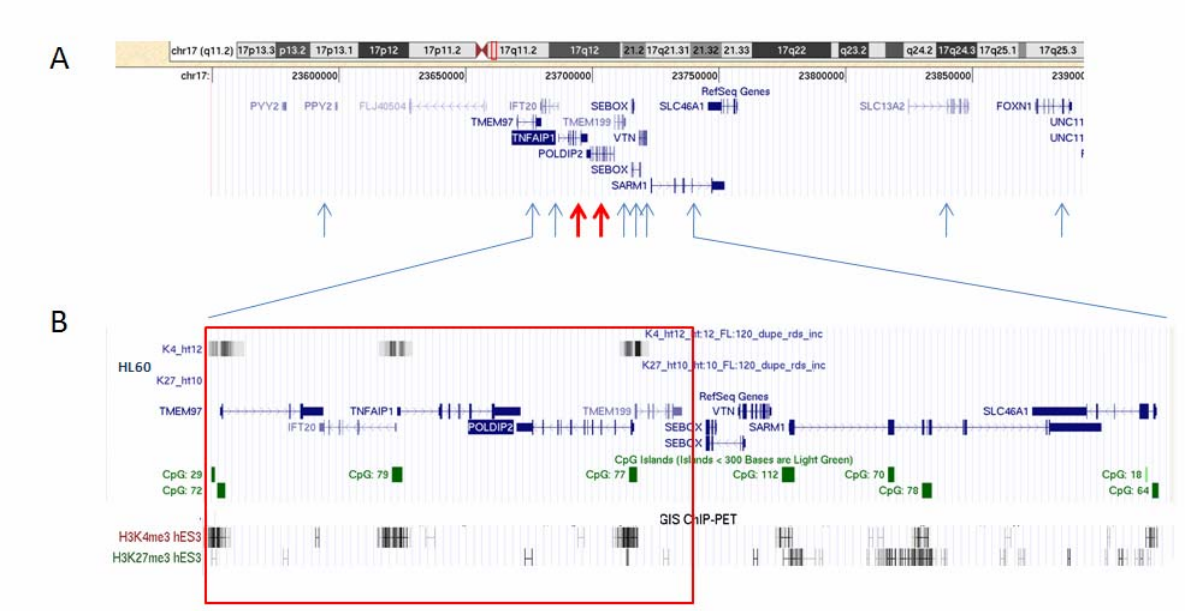

Fig. 2 TNFAIP1/POLDIP2 is complex cis-sense antisense gene architecture. A - TNFAIP1/POLDIP cis-sense antisense gene pair (thick arrows) with 9 neighboring genes included into the analysis (thin arrows). B - TNFAIP1/POLDIP2 complex cis-sense antisense gene architecture (rectangle) including tracks for DNA association with triplemethylated histone H3K4me3 and H3K27me3. Triplemethylated histone H3K4me3 and H3K27me3 are reproducibly occurred in different types of cell lines (HL60 (Chip-seq analysis [37]) and hES3 (Chip-PET analysis [38])); H3K4me3 is associated with active/open chromatin regions, H3K27me3 is associated with repressed/closed chromatin regions, however, H3K27me3 does not found in TNFAIP1/POLDIP2.

TABLE I P-VALUES OBTAINED BY PAIR-WISE COMPARISONS OF MATRICES FOR 5 GENES IN SFGM GROUP (SFGMM - SFGM MATRIX) AND 6 "NEIGHBORING" GENES (NM – "NEIGHBORS" MATRIX). [1] - P-VALUES WERE CALCULATED USING BOOTSTRAP BARTLETT'S TEST; [2] - P-VALUES WERE CALCULATED USING BOX'S M TEST (SEE DESCRIPTION OF PROCEDURES IN MATERIALS AND METHODS SECTION); IGM – INTERGENIC MATRIX; U – UPPSALA COHORT, S – STOCKHOLM COHORT (SEE METHODS)

Breast cancer grade	SFGMM [1]		NM [1]		SFGMM/NM [2]		IGM/NM [2]	
	U	S	U	S	U	S	U	S
G3	1.0E-11	2.0E-03	8.1E-01	5.0E-01	1.0E-16	1.0E-16	2.4E-01	4.5E-01
G3-like	1.0E-11	1.90E-02	5.00E-01	9.8E-01	1.0E-16	3.4E-01	2.3E-02	3.3E-01
G1-like	1.0E-14	1.2E-01	3.9E-01	5.0E-01	1.0E-16	1.0E-16	2.7E-01	5.1E-01
G1	1.0E-04	1.0E-15	9.5E-01	7.2E-01	1.0E-16	1.0E-16	9.9E-02	1.8E-01
Total group	1.0E-16	1.0E-15	7.3E-01	6.6E-01	1.0E-16	1.0E-15	4.0E-01	6.9E-01

Analysis of correlation matrices of these 11 genes identified a phenomenon when 5 genes structurally organized as CSAGA and tightly linked in the genome (TMEM97/IFT20/ TNFAIP1 /POLDIP2 /TMEM199, Figure 2, B) showed a strong coregulatory pattern in breast cancer patients (Figure 3, A, B) in Uppsala as well as in Stockholm breast cancer cohorts. Moreover, expression level for each of the 5 genes in different grades in both cohorts was much stronger comparing to their neighbors in the chosen genomic window (Figure 3, C and D). At the same time, moderate significant differences in genes expression level were observed for TMEM97 and POLDIP2 genes in different grades in both breast cancer cohorts (not shown).

A structural backbone of this TNFAIP1/POLDIP2 CSAGA is composed of 3 CpG rich regions representing putative gene promoters (2 of which are bidirectional), as well as of 2 intergenic convergent SA overlaps (TMEM97 vs IFT20, TNFAIP1vs POLDIP2) with RefSeq support (Figure 2, B) and

1 divergent SA overlap with UCSC support (IFT20 vs TNFAIP1) (not shown).

Based on the observation of its structural and expressional integrity, we termed TNFAIP1/POLDIP2 CSAGA as a TNFAIP1/POLDIP2 structural-functional gene module (SFGM). For the remaining 6 genes in the chosen window we proposed the term "neighboring" genes for further description convenience.

Next, using statistical procedures based on Bartlett' [25] and Box' M tests[27], we addressed the following questions: 1) whether the correlation matrices for the 5 genes of TNFAIP1/POLDIP2 SFGM (SFGM matrix -SFGMM) as well as correlation matrices for the 6 "neighboring" genes ("neighbors" matrix – NM) are statistically significant comparing with randomly chosen matrices in the whole genome (Figure 3, A and B); 2) whether SFGMMs are statistically significant from NMs; 3) whether intergenic matrices (IGMs) are not statistically different from NMs. The results are presented in Table 1. In Uppsala cohort, correlation matrices were highly significant in total group as well as in 4 different grades. In Stockholm cohort, matrices for total group, G1 and G3 subgroups were highly significant. G3-like subgroup was close to the border line ($p<0.01$) and only G1-like was not significant (Table 1). Due to the fact that for all NMs in both cohorts p-values in Barlett's test were at least several times higher than the borderline, we excluded them as candidates for the members of TNFAIP1/POLDIP2 SFGM.

We also applied Box' M test for comparing two correlation matrices at α= 1%. The test revealed highly significant differences in almost all matrices pairs SFGMs vs NMs (except Stockholm G3-like) and absence of differences in all matrices pairs IGMs vs NMs. Taken together, the statistical analysis clearly supports the existence of the 5 gene's structural and functional gene module. On the other hand, it strongly excludes the 6 studied "neighboring" genes as members of this SFGM. Noteworthy, we also utilized Box's M test to compare if there are any differences among SFGM matrices for each grade inside of both of the cohorts. Surprisingly, we observed that SFGM showed significant strengthening of coregulatory profile (from left to right) in the following group pairs: G1 vs G3-like (Uppsala, $p<8.08E-03$, Stockholm, $p<9.21E-07$), G1-like vs G3 (Stockholm, $p<3.46E-004$), G1 vs G3 (Stockholm, $p<2.62E-04$), G1-like vs G3-like (Stockholm, $p<3.64E-08$).

Survival analysis of SFGM genes and their closest neighbors in breast cancer patients

Survival analysis algorithm and the software were developed in our group previously [28].

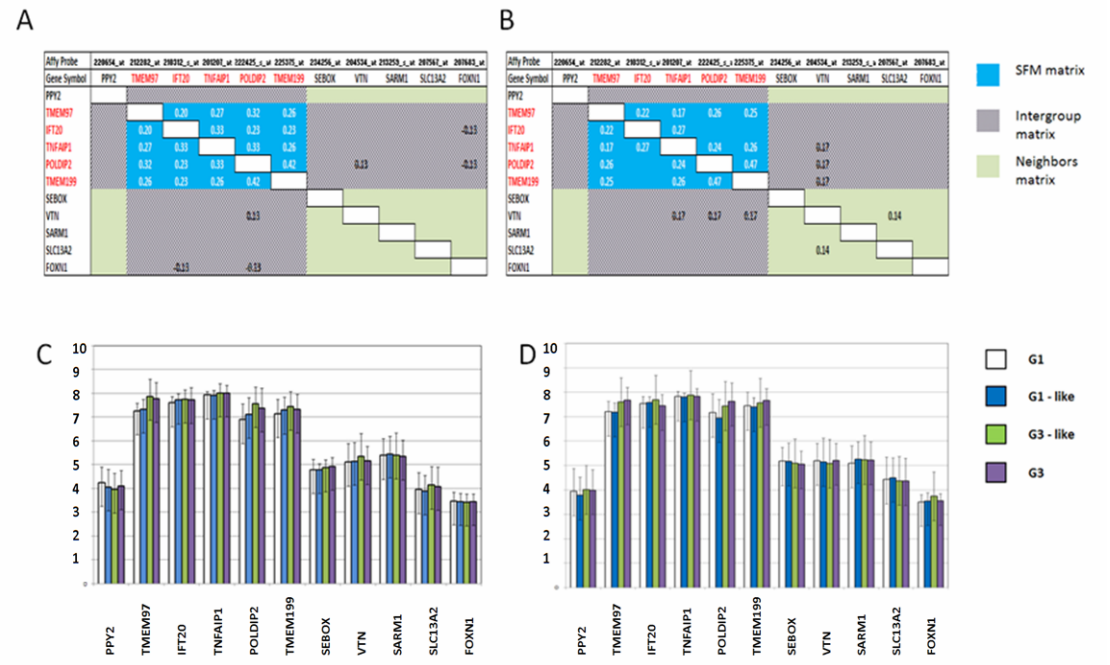

Fig. 3. Members of the TNFAIP1/POLDIP2 are co-expressed in breast cancer and organized in the structural –functional gene module (SFGM). Visualization of correlation matrices of expression of CSAGA genes demonstrates the presence of strong co-regulatory pattern: TNFAIP1/POLDIP2 co-regulatory area is defined by enrichment of significant correlation coefficients (Z-values, $p<0.01$). Members of TNFAIP1/POLDIP2 form the SFGM (*SFGM matrix -SFGMM*). *Neighbor's matrix* shows correlations among 6 "neighboring" genes (*Neighbor's matrix* – NM). *Intergroup matrix* provides intergroup correlations between genes of SFGM and NM. **A, C** – Uppsala cohort; **B, D** – Stockholm cohort; A and **B** – correlation matrices, **C** and **D** - genes expression data (Mean+/-SD) in breast cancer patients with different genetic grades (G1, G1-like, G3-like, G3, [24]).

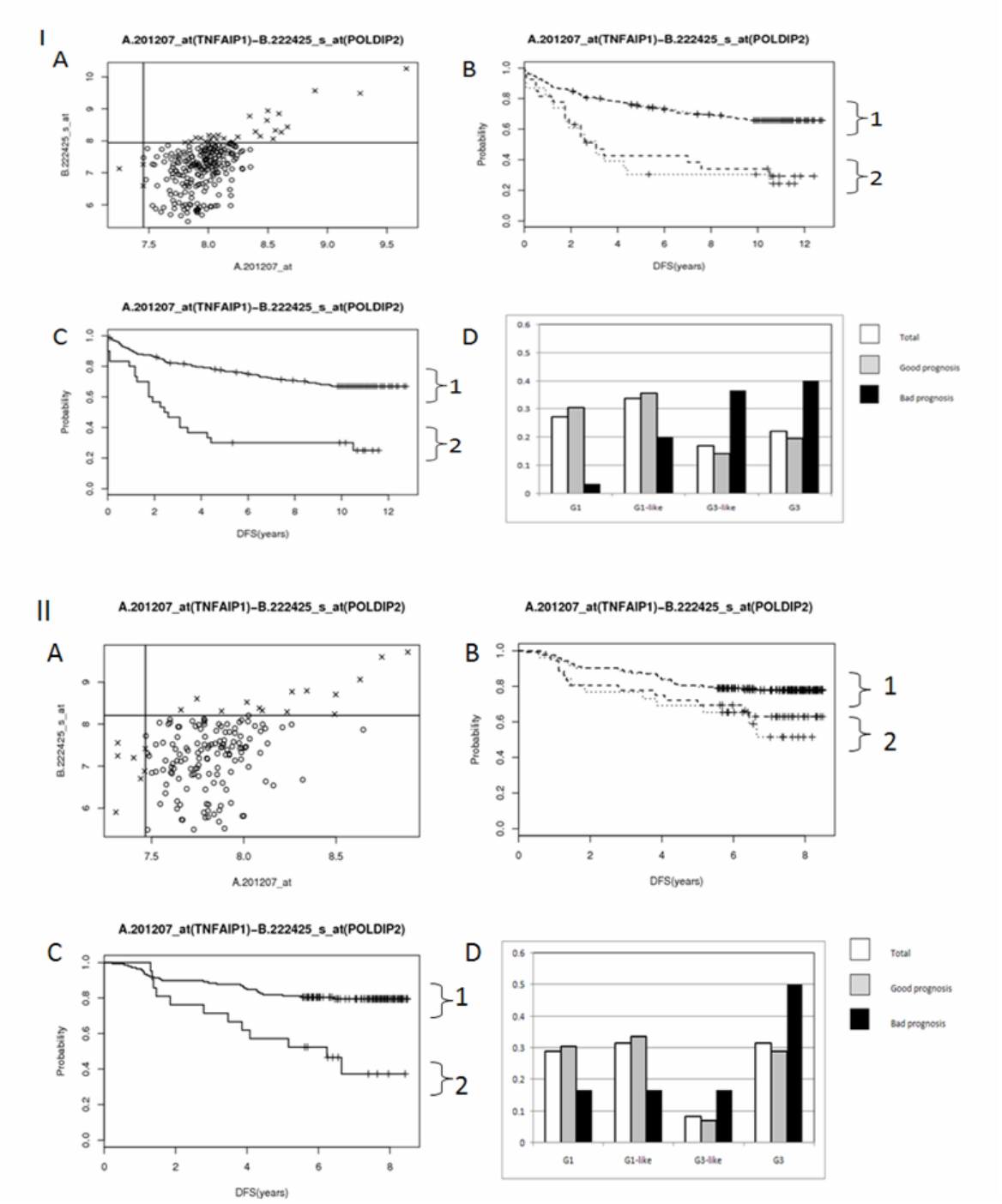

Fig. 4. Survival analysis for the TNFAIP1/POLDIP2 gene pair in breast cancer patients: I -Uppsala cohort, II – Stockholm cohort. A - Correlation of the gene expression for TNFAIP1 and POLDIP2 genes and optimal partition of expression domains; B – survival curves for individual genes (when analyzed separately): dashed - TNFAIP1, dotted - POLDIP2; C – survival curves for the gene pair TNFAIP1/POLDIP2 (when analyzed together); D - distribution of patients in different grades (white column - total group, grey column – good prognosis group, black column – bad prognosis group). For B and C – survival curve(s) **1** correspond to good prognosis group, survival curve(s), **2** correspond to bad prognosis group.

TABLE II INDIVIDUAL GENES SELECTED AMONG TNFAIP1/POLDIP2 SFGM AND 6 "NEIGHBORING" GENES WHICH PROVED TO BE SURVIVAL SIGNIFICANT IN BOTH UPPSALA AND STOCKHOLM COHORTS (P-VALUE ≤ 0.05). UPPSALA P-VALUE CORRECTION P*U=4.2E-03; STOCKHOLM P-VALUE CORRECTION P*U=5.1E-03.

Affymetrix U133 (A&B) probe	Gene Symbol	Uppsala cohort		Stockholm p-value (individual)	
		Wald statistic p-value	FDR corrected p-value = 4.2E-03	Wald statistic p-value	FDR corrected p-value = 5.1E-03
B.222425_s_at	POLDIP2	1.5E-05	Significant	2.4E-02	Not significant
A.210312_s_at	IFT20	4.1E-04	Significant	2.4E-02	Not significant
A.212282_at	TMEM97	1.1E-03	Significant	3.0E-03	Significant
B.225375_at	TMEM199	2.2E-02	Not significant	7.2E-04	Significant

We isolated individually survival significant genes (Table 2). As it is evident from Table 2, four members (unique genes) of TNFAIP1/POLDIP2 SFGM are significant at α=5%, whereas

no neighboring genes satisfied this criterion. To minimize Type I error rate (false positives) we applied False Discovery Rate (FDR) correction to the p-values using the classic FDR of Benjamini and Hochberg, 1995[34], extended for positive dependent data [36]. Typically, positive dependent exists if the variance covariance matrix of the six probes we study contains only positive entries, which is true in our case. At significance level α = 5%, the Uppsala and the Stockholm p-value corrections were estimated to be p^*_u=4.2E-03 and p^*_u=5.1E-03, respectively. Table 2 indicates the Wald and FDR significant probes of our set. Notice that after FDR correction our set still contains highly significant genes in both cohorts. From Table 2 we can observe that only genes belonging to the TNFAIP1/POLDIP2 SFGM are survival significant at least in one cohort; TMEM97 was survival significant in both cohorts, at the same time it was shown previously to play a role in primary and metastatic colorectal cancers [36]. We also applied survival analysis and 2D DDg to identify survival significant probe pairs among the probes of our prospective cluster. Analysis was performed in such a way that the designs (e.g., Figure 4, I-A, II-A) across the cohorts were the same (patients were similarly distributed). First, we estimated the Wald p-values and then used the FDR correction as before. The corrected p-values in Uppsala and Stockholm cohorts were $p^*_u = 8.5E-03$ and $p^*_s = 4.9E-03$. We isolated those gene pairs which were survival significant in both cohorts studied. Table 3 shows the 10 non-redundant survival significant gene pairs identified in our analysis.

Among the eight different genes from significant gene pairs in Table 3 we observed 2 (SARM and VTN) belonging to the "neighboring" genes of TNFAIP1/POLDIP2 SFGM. This fact may be explained that on the level of clinical phenotype gene-gene interactions of TNFAIP1/POLDIP2 SFGM genes with their "neighbors" could be much more complex, than on the transcriptional level. At the same time we could speculate that TNFAIP1/POLDIP2 SFGM may serve as a local "core" region of survival significant and clinically important genes which may "drive" the surrounding genes in breast cancer progression.

Noteworthy, 3 out of 10 isolated gene pairs (bold italic in Table 3) demonstrated strong effect of synergy. Wald p-value calculated for a gene pair (when expressional and clinical data were analyzed for both genes) was at least 10 times lower comparing to p-values calculated for individual genes of the pair.

TNFAIP1/POLDIP2 gene pair showing a strong synergy effect on survival (more than 20 times in both cohorts, Table 3) in fact is a convergent SA pair in the middle of TNFAIP1/POLDIP2 SFGM. Graphical representation of survival significant genes pair TNFAIP1/POLDIP2 with more details is shown in Figure 4.

Therefore, survival analysis of TNFAIP1/POLDIP2 SFGM and its "genes-neighbors" revealed individual survival significant genes as well as significant gene pairs, and this fact may suggest an importance of TNFAIP1/POLDIP2 SFGM for breast cancer prognosis.

TABLE III SELECTED NON-REDUNDANT SURVIVAL SIGNIFICANT GENE PAIRS IDENTIFIED IN BOTH COHORTS OF BREAST CANCER PATIENTS. AFFY PROBE* : AFFYMETRIX U133 (A&B) PROBES IDS, GS*: GENE SYMBOL. U: UPPSALA COHORT; S: STOCKHOLM COHORT

Affyprobe*	GS*	p-value (individual)		Affyprobe*	GS*	p-value(individual)		p-value(gene pair)	
		U	S			U	S	U	S
B.222425_s_at	***POLDIP2***	***1.50E-05***	***0.024***	***A.201207_at***	***TNFAIP1***	***0.00011***	***0.081***	***3.10E-07***	***0.00046***
A.201208_s_at	***TNFAIP1***	***0.022***	***0.11***	***A.214283_at***	***TMEM97***	***0.074***	***0.081***	***0.00022***	***0.0029***
A.213259_s_at	***SARM1***	***0.011***	***0.074***	***B.225375_at***	***TMEM199***	***0.022***	***0.00072***	***0.00085***	***2.90E-05***
A.204534_at	VTN	0.024	0.11	A.210312_s_at	IFT20	0.00041	0.024	0.00021	0.00052
A.204534_at	VTN	0.024	0.11	A.212279_at	TMEM97	0.0042	0.0035	1.00E-04	0.00062
A.210312_s_at	IFT20	0.00041	0.024	A.212281_s_at	TMEM97	0.0028	0.0051	1.60E-05	0.0036
A.213259_s_at	SARM1	0.011	0.074	A.212281_s_at	TMEM97	0.0028	0.0051	0.00039	0.004
A.217806_s_at	POLDIP2	4.30E-05	0.12	A.212281_s_at	TMEM97	0.0028	0.0051	2.60E-05	0.0036
B.225375_at	TMEM199	0.022	0.00072	A.201207_at	TNFAIP1	0.00011	0.081	9.30E-05	0.00021
B.225375_at	TMEM199	0.022	0.00072	A.212279_at	TMEM97	0.0042	0.0035	2.00E-04	0.00032

Genes of TNFAIP1/POLDIP2 SFGM could be coregulated by chromatin remodeling/activation.

Figure 2, B demonstrates several facts which could indicate histones modification as possible mechanism of observed transcriptional coregulatory pattern of TNFAIP1/POLDIP2 SFGM. Additional customer tracks in UCSC browser for trimethylated histones H3K4me3 and H3K27me3 DNA association in promyelocytic leukemia cells (HL60) [37], (http://www.bcgsc.ca/data/histone-modification) confirmed the fact of transcriptional activation of the genes involved in the TNFAIP1/POLDIP2 SFGM. All three CpG–rich putative promoters in the TNFAIP1/POLDIP2 SFGM showed clear binding with H3K4me3 (a marker of transcriptional active chromatin) as well as lack of binding with H3K27me3 (a marker for inactive chromatin). Nevertheless, putative promoters for the "neighboring" genes SEBOX, VTN, and SARM did not show any binding signal for H3K4me3. Similar situation is observed for GIS Chip-PET track (embryonic stem cells hES3) of UCSC Browser [38] (Figure 2, B). Strong DNA binding signal of H3K4me3 is observed in all three putative promoter regions of TNFAIP1/POLDIP2 SFGM and only faint binding of H3K27me3 is visible for TMEM97 and TMEM199/POLDIP2 putative promoters. Alternatively, putative promoter for SEBOX gene does not show any signal for both H3K4me3 and H3K27me3; regulatory region for VTN and SARM1 reveals moderate signal for H3K4me3 and H3K27me3 of the same intensity. Although the mentioned tracks show DNA binding data not for breast cancer cell lines, we would not exclude the possibility of similar mechanism of coregulatory pattern of TNFAIP1/POLDIP2 SFGM also in breast cancer cell lines and patients.

IV. Discussion

A method of statistical identification of co-regulated genes organized in complex genome architectures

In the present study, we have developed a new computational method of statistical identification of co-regulated genes organized in complex genome architectures including more than one SA gene pair. Our approach is based on: (i) concordant analysis and selection of expressed SA genes, (ii) identification of the boundaries of a genomic region encompassing genes with similar co-expression pattern, (iii) validation of the stability of the expression pattern using independent patient cohorts, (iv) evaluation of clinical significance of expressed genes which belong to identified genome region and (v) identification of synergy of the genes in context of disease aggressiveness and disease relapse.

TNFAIP1/POLDIP2 is an essential structural-functional module in the human genome

In present work we performed the analysis of the TNFAIP1/POLDIP2 CSAGA on 17q11.2 in two breast cancer cohorts as an example of implementation of our approach. TNFAIP1/POLDIP2 CSAGA is composed of 5 genes: TMEM97, IFT20, TNFAIP1, POLDIP2 and TMEM199. The gene pairs (TMEM97 - IFT20), (TNFAIP1- POLDIP2) and (IFT20-TNFAIP1) produce SA transcripts; gene pairs (IFT20-TNFAIP1) and (POLDIP2 - TMEM199) share corresponding bi-directional promoter regions. This complex genomic region exhibits a well-organized transcription apparatus : 3 CpG islands; several TF binding sites in canonical promoter regions – GATA1, TAXCREB, CREBP1, CREB, SREBP1 (http://genome.ucsc.edu/cgi-bin/hgTracks, Transfac 7.0); strong signals for POL-II binding (not shown) and probable open chromatin regions (H3K4met3(+) and H3K27met3(-) regions (Figure 1, B)). TNFAIP1/POLDIP2 CSAGA region could produce a large diversity of alternative splice variants of the genes and is highly enriched with many other regulatory sequences (USCS genome browser, AceView Gene Models with Alternative Splicing). Analysis of correlation matrices revealed a phenomenon when genes structurally organized in the genome in CSAGA demonstrate reproducible coregulatory pattern in breast cancer cells (Figure 3). We termed TNFAIP1/POLDIP2 CSAGA as the TNFAIP1/POLDIP2 structural-functional gene module (SFGM).

Concordant positive regulation in TNFAIP1/POLDIP2 CSAGA

We did not observe any significant negative correlations (discordant regulation) in the TNFAIP1/POLDIP2 SFGM what is in agreement with several previous reports of frequent concordant regulation of sense-antisense pairs [33], [39]-[41].

Correlation matrices analysis of TNFAIP1/POLDIP2 SFGM in four grades (G1, G1-like, G3-like and G3) of breast cancer patients identified the fact of strengthening of correlations between the genes of TNFAIP1/POLDIP2 SFGM. This finding indicates the importance of the module in breast cancer progression. On the other hand, this idea is supported by survival analysis of individual genes as well as of gene pairs of TNFAIP1/POLDIP2 SFGM and its neighbors. Only the genes of TNFAIP1/POLDIP2 SFGM proved to be survival significant in at least 1 of 2 cohorts analyzed. Among 11 genes analyzed, 10 survival significant gene pairs have been identified and all the genes of the TNFAIP1/POLDIP2 SFGM were involved in these pairs. Each of the 10 pairs contained at least 1 gene of the SFGM. Moreover, 3 top level survival significant gene pairs demonstrated strong synergetic effect in prognosis of breast cancer disease relapse when compared with individual genes.

Protein interaction sub-network

Our analysis of the literature about members of the module confirmed our suggestion regarding functional integrity of TNFAIP1/POLDIP2 SFGM and its importance in cancers.

Liu et al. (2003) [42] reported about physical interaction of POLDIP2 protein with p50 subunit of DNA polymerase delta and proliferating cell nuclear antigen (PCNA). PCNA has been called the "ringmaster of the genome", because it has been shown to actively participate in a number of the molecular pathways responsible for the life and death of the mammalian cell [43]. It also has proven to be a useful marker to evaluate cell proliferation and prognosis when combined with other breast cancer markers, such as estrogen receptor, progesterone receptor and ERBB2 [44]-[46]. Another work suggested that rat TNFAIP1 is homologous to polymerase delta-interacting protein (PDIP1) as well as with PCNA. Both rat PDIP1 and rat TNFAIP1 could stimulate polymerase delta activity in vitro in PCNA-dependent way [47]. TMEM97 cytoplasmic expression was shown to be positively related to

expression of PCNA and considered to be as a prognostic factor in metastasis of colorectal cancer [36]. Therefore, at least 3 members of TNFAIP1/POLDIP2 module could be functionally associated in the same PCNA complex.

Two interesting recent publications support the idea about involving TNFAIP1/POLDIP2 module in cell cycle and cells proliferation. POLDIP2 was shown to be associated with spindle organization and aberrant chromosome segregation [48]. Tissue specific floxed deletion of IFT20 in the mouse kidney causes mis-orientation of the mitotic spindle in collecting duct cells, prevents cilia formation and promotes rapid postnatal cystic expansion of the kidney [49].

Interesting pleiotropic effects of POLDIP2 also include interaction with cell-cell adhesion receptor CEACAM1 [50] and involvement in transcription and metabolism of mitochondrial DNA [51].

It is important to note that TNFAIP1/POLDIP2 module is located outside of the well-known ERBB2 amplicon on 17q12, overrepresentation of which in the genome often associated with occurrence of ERBB2-positive breast cancer subtype. In the future studies it is intriguing to analyze the relationships of TNFAIP1/POLDIP2 module and ERBB2 amplicon region.

Taken together, our analysis suggests that TNFAIP1/POLDIP2 SFGM is composed of the genes which are not only closely organized in a complex genomic architecture and co-regulated on the epigenetic and transcription levels, but which are also could be involved in essential biochemical pathways as well as protein-protein interactions forming molecular complexes important for many cell needs, including cell division, proliferation, apoptosis, intracellular transport and cell binding. Such non-random combinations of structural and functional properties of the gene architectures suggest evolution and physiological essentiality and clinical significance of the TNFAIP1/POLDIP2. Transcription co-activation of genes in this CSAGA is strongly associated with high aggressiveness and poor prognosis of breast cancer.

V. Conclusion

We can conclude that the methods of computational identification of novel structural-functional gene modules and data-driven grouping of clinically heterogeneous (cancer) patients based on expression patterns of genes of such modules could provide broad perspectives of development of computational systems biology strategies for understanding genetics and pathobiology of complex genetic diseases.

Transcription co-activation of genes in TNFAIP1/POLDIP2 is strongly associated with this CSAGA amplification, high aggressiveness and poor prognosis of breast cancer. Due to concordant regulation pattern of genes in structural-functional gene modules, one could either target the antisense transcript(s) along, resulting in reduction of expression of sense mRNA transcription from sense gene and even adjacent genes of the SA pair. Due to such possibility, pharmacological strategies aimed at either stimulation or suppression of expression profile for a specific group of genes which are influenced by natural SA regulation could be also developed. A discovery of biologically meaningful and clinically significant CSAGAs instead of conventional finding of "gene signatures" might be more promising in context of (i) understanding of the mechanistic role of CSAGAs in complex diseases including cancer, (ii) efficiency of microarray analysis into clinical practice, (iii) identification of new drug targets and (iv) development of new drug strategies.

References

[1] V.A. Kuznetsov, G.D. Knott, R.F. Bonner, "General statistics of stochastic process of gene expression in eukaryotic cells," *Genetics,* vol.161(3), pp.1321-32, 2002.

[2] Y. Zhang, X.S. Liu, Q.R. Liu and L .Wei, "Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species," Nucleic Acids Res, vol. 34, pp. 3465-75, 2006.

[3] P.G. Engström, H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa et al., "Complex loci in human and mouse genomes,". *PLoS Genet.,* vol.2(4), pp.e47, 2006.

[4] P. Kapranov, A.T. Willigham, T.R. Gingeras, "Genome-wide transcription and the implications for genomic organization," *Nat Rev Genet*, vol.8(6), pp.413-423, 2007.

[5] I.B. Rogozin, A.N. Spiridonov, A.V. Sorokin, Y.I. Wolf, I.K. Jordan et al., "Purifying and directional selection in overlapping prokaryotic genes," *Trends Genet,* vol.18(5), pp. 228-322002.

[6] Z.I. Johnson and S.W. Chisholm, "Properties of overlapping genes are conserved across microbial genomes," *Genome Res.*, vol.14(11), pp. 2268-72, 2004.

[7] E. Enerly, Z. Sheng and K.B. Li, "Natural antisense as potential regulator of alternative initiation, splicing and termination," *In Silico Biol,* vol.5(4), pp. 367-77, 2005.

[8] C.M. Henderson, C.B. Anderson, M.T. Howard, "Antisense-induced ribosomal frameshifting," *Nucleic Acids Res*, vol.34(15), pp. 4302-10, 2006.

[9] U. Orfanelli, A.K. Wenke, C. Doglioni, V. Russo, A.K. Bosserhoff, G. Lavorgna, "Identification of novel sense and antisense transcription at the TRPM2 locus in cancer," *Cell Res*, vol.18, pp. 1128-1140, 2008.

[10] E. Gallagher, A. Mc Goldrick, W.Y. Chung, O. Mc Cormack, M. Harrison et al., "Gain of imprinting of SLC22A18 sense and antisense transcripts in human breast cancer," *Genomics*, vol. 88(1), pp.12-7, 2006.

[11] W. Yu, D. Gius, P. Onyango, K. Muldoon-Jacobs, J. Karp et al., "Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA, " *Nature,* , vol.451 (7175), pp. 202-6, 2008.

[12] Y. Ogawa and J.T. Lee, "Antisense regulation in X inactivation and autosomal imprinting," *Cytogenet. Genome Res*., vol.99(1-4), pp.59-65, 2002.

[13] G. Alfano, C. Vitiello, C. Caccioppoli, T. Caramico, A. Carola et al., "Natural antisense transcripts associated with genes involved in eye development, " *Hum Mol Genet*, vol.14(7), pp. 913-23, 2005.

[14] J.H. Guo, H.P. Cheng, L. Yu and S. Zhao, "Natural antisense transcripts of Alzheimer's disease associated genes, " *DNA Seq*, vol.17(2), pp.170-3, 2006.

[15] P.Juszczynski, J.L. Kutok, C. Li, J. Mitra, R.C. Aguiar, M.A. Shipp,"BAL1 and BBAP are regulated by a gamma interferon-responsive bidirectional promoter and are overexpressed in diffuse large B-cell lymphomas with a prominent inflammatory infiltrate," *Mol Cell Biol.,* vol.26(14), pp.5348-59, 2006.

[16] Kuznetsov V.A. *et al.* (2006) Genome-wide co-expression patterns of human cis-antisense gene pairs," *Proc. of the 5-th Intern Conf on Bioinformatics of Genome Regulation and Structures.* Novosibirsk, Inst. of Cytology&Genetics, 1, 90-93.

[17] Y. Ohinata, S. Sutou, M. Kondo, T. Takahashi, Y. Mitsui, " Male-enhanced antigen-1 gene flanked by two overlapping genes is expressed in late spermatogenesis," *Biol Reprod.* Vol.67(6), pp.1824-31, 2002.

[18] V. Veeramachaneni, W. Makalowski, M. Galdzicki, R. Sood and I. Makalowska, "Mammalian overlapping genes: the comparative perspective," *Genome Res*, vol.14(2), pp.280-6, 2004.

[19] I. Makalowska, C.F. Lin and W. Makalowski, "Overlapping genes in vertebrate genomes," *Comput Biol Chem*, vol.29(1), pp. 1-12. 2005.

[20] Hu, H.M. Stern, L. Ge, C. O'Brien, L. Haydu, " Genetic alterations and oncogenic pathways associated with breast cancer subtypes," *Mol Cancer Res*., vol.7(4), pp.511-22, 2009.

[21] P.M. Haverty, J. Fridlyand, L. Li, G. Getz, R. Beroukhim et al., "High-resolution genomic and expression analyses of copy number alterations

in breast tumors," *Genes Chromosomes Cancer.*, vol.47(6), pp.530-42, 2008.

[22] I .Vanden Bempt, M. Drijkoningen, C. De Wolf-Peeters, " The complexity of genotypic alterations underlying HER2-positive breast cancer: an explanation for its clinical heterogeneity," *Curr Opin Oncol*. vol.19(6), pp. 552-7, 2007.

[23] Rody A, Karn T, Ruckhäberle E, Müller V, Gehrmann M, Solbach C, Ahr A, Gätje R, Holtrich U, Kaufmann M. Gene expression of topoisomerase II alpha (TOP2A) by microarray analysis is highly prognostic in estrogen receptor (ER) positive breast cancer. Breast Cancer Res Treat. 2009 Feb;113(3):457-66

[24] A.V. Ivshina, J. George, O. Senko, B. Mow, T.C. Putti et al., "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer Res.,* vol.66(21), pp. 10292-301, 2006.

[25] M.S. Bartlett, "Test of significance in factor analysis," *Br. J. Psychol.*, vol.3, pp.77 – 85, 1950.

[26] B. Efron, and R.J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1994.

[27] G. E. P. Box," A general distribution theory for a class of likelihood criteria", *Biometrika*, vol.36, pp. 317–346, 1949.

[28] E. Motakis, A.V. Ivshina, V.A. Kuznetsov, "Identification of essential genes and gene pairs associated with survival time of cancer patients**".** In *Proceedings of the 2007 international conference on bioinformatics& computational biology (BIOCOMP 2007): 25-28 June 2007; Las Vegas Nevada USA.* Edited by Hamid R. Arabnia, Mary Qu Yang and Jack Y. Yang: CSREA Press; vol.II: pp.753-759, 2007.

[29] R.D. Cox and D. Oakes, *Analysis of Survival Data*. London: Chapman and Hall, 1984.

[30] Y.L. Orlov, J. Zhou, L. Lipovich, A. Shahab, V.A. Kuznetsov, "Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis," *In Silico Biol.*, vol.7(3), pp.241-60, 2007.

[31] E. Arriola, C. Marchio, D.S. Tan, S.C. Drury, M.B. Lambros, R. Natrajan, S.M. Rodriguez-Pinilla, A. Mackay, N. Tamber, K. Fenwick et al., "Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines", *Lab Invest,* vol.88(5), pp.491-503, 2008.

[32] P. Deloukas, G.D. Schuler, G. Gyapay, E.M. Beasley, C. Soderlund et al. "A Physical Map of 30,000 Human Genes", *Science*, vol.282(5389), p.744 – 746, 1998.

[33] V.A. Kuznetsov, J.T. Zhou, J. George and Yu.L. Orlov, "Genome-wide co-expression patterns of human cis-antisense gene pairs". In *Proc of the 5-th Intern Conf on Bioinformatics of Genome Regulation and Structures: Novosibirsk, Inst. of Cytology&Genetics,* vol.1, pp. 90-93, 2006.

[34] Y. Benjamini and Y.Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society*. Series B (Methodological), vol.57, pp.289–300, 1995.

[35] Y. Benjamini and Yekutieli D., "The control of the false discovery rate in multiple testing under dependency", *Annals of Statistics*, vol.29, pp. 1165–1188, 2001.

[36] S.B. Moparthi, G. Arbman, A. Wallin, H. Kayed, J. Kleeff et al., "Expression of MAC30 protein is related to survival and biological variables in primary and metastatic colorectal cancers," *Int J Oncol.,* vol.30(1), pp.91-5, 2007.

[37] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones et al. "High-resolution profiling of histone methylations in the human genome" *Cell*, vol.129(4), pp. 823-37, 2007.

[38] X.D. Zhao, X. Han, J.L. Chew, J. Liu, K.P. Chiu et al., "Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells". *Cell Stem Cell*, vol.1(3), pp.286-98,2007.

[39] R. Yelin, D. Dahary, R. Sorek, E.Y. Levanon, O. Goldstein et al. "A Widespread occurrence of antisense transcription in the human genome", *Nat Biotechnol*, vol.21, pp. 379-86, 2003.

[40] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi et al. "Antisense transcription in the mammalian transcriptome," *Science,* vol.309(5740), pp.1564-6, 2005.

[41] M. Baguma-Nibasheka, A.W. Li, M.S. Osman, L. Geldenhuys, A.G. Casson et al., "Coexpression and regulation of the FGF-2 and FGF antisense genes in leukemic cells", *Leuk Res*., vol.29(4), pp.423-33.

[42] L. Liu, E.M. Rodriguez-Belmonte, N. Mazloum, B. Xie and M.Y. Lee, "Identification of a novel protein, PDIP38, that interacts with the p50 subunit of DNA polymerase delta and proliferating cell nuclear antigen," *J.Biol Chem*., vol. 278(12), pp.10041-7, 2003.

[43] T. Paunesku, S. Mittal, M. Protić, J. Oryhon, S.V. Korolev, A. et al., "Proliferating cell nuclear antigen (PCNA): ringmaster of the genome", *Int. J. Radiat. Biol.*, vol.77 (10), pp.1007-21, 2001.

[44] S. Aaltomaa, P. Lipponen, K. Syrjänen, "Proliferating cell nuclear antigen (PCNA) immunolabeling as a prognostic factor in axillary lymph node negative breast cancer,"Anticancer Res., vol. 13(2), pp.533-8, 1993.

[45] G.W. Jr. Sledge, K.D. Miller, " Exploiting the hallmarks of cancer: the future conquest of breast cancer," *Eur. J. Cancer.,* vol. 39(12), pp.1668-75, 2003.

[46] L.H. Malkas, B.S. Herbert, W. Abdel-Aziz, L.E. Dobrolecki, Y. Liu, et al., "A cancer-associated PCNA expressed in breast cancer has implications as a potential biomarker," *Proc Natl Acad Sci U S A*, vol.103(51), pp.19472-7, 2006.

[47] J. Zhou, X. Hu, X. Xiong, X. Liu, Y. Liu et al., "Cloning of two rat PDIP1 related genes and their interactions with proliferating cell nuclear antigen," *J Exp Zoolog A Comp Exp Biol,* vol.303(3), pp.227-40, 2005.

[48] E. Klaile, A. Kukalev, B. Obrink, M.M. Muller, "PDIP38 is a novel mitotic spindle-associated protein that affects spindle organization and chromosome segregation", *Cell cycle*, vol. 7(20), pp. 3180-3186, 2008.

[49] J.A. Jonassen, J. San Agustin, J.A. Follit, G.J. Pazour, "Deletion of IFT20 in the mouse kidney causes misorientation of the mitotic spindle and cystic kidney disease," *J. Cell Biol.*, vol. 183(3), pp.377-84, 2008.

[50] E. Klaile, M.M. Müller, C. Kannicht, W. Otto, B.B. Singer et al., "The cell adhesion receptor carcinoembryonic antigen-related cell adhesion molecule 1 regulates nucleocytoplasmic trafficking of DNA polymerase delta-interacting protein 38," *J. Biol. Chem.,* vol. 282(36), pp.26629-40, 2007.

[51] X. Cheng, T. Kanki, A. Fukuoh, K. Ohgaki, R. Takeya et al., "PDIP38 associates with proteins constituting the mitochondrial DNA nucleoid," *J. Biochem.*, vol.138(6), pp.673-8, 2005.

[52] F.W. Wolf, R.M. Marks, V. Sarma, M.G. Byers, R.W. Katz et al., "Characterization of a novel tumor necrosis factor-alpha-induced endothelial primary response gene," J. Biol. Chem., vol.267(2), pp. 1317-26, 1992.