A Transcriptional Study of Oncogenes and Tumor Suppressors Altered by Copy Number Variations in Ovarian Cancer

Giorgia Giacomini[1,3,4], Gabriele Ciravegna[2,4], Marco Pellegrini[3], Romina D'Aurizio[3], and Monica Bianchini[4]

[1] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Italy
[2] Department of Information Engineering, University of Florence, Italy
[3] Institute of Informatics and Telematics, CNR, Pisa, Italy
[4] Department of Information Engineering and Mathematics, University of Siena, Italy
giacomini@student.unisi.it gabriele.ciravegna@unifi.it

Abstract. The most popular approach to explain cancer is based on the discovery of oncogenes and tumour suppressor genes as a preliminary step in estimating their impact on altered pathways. The present paper proposes a pipeline which aims at detecting 'weak' or 'indirect' functions impacted by Copy Number Variations (CNVs) of cancer–related genes, integrating such signals over all known oncogenes/tumour suppressor genes of a cancer type. We applied the pipeline to the task of detecting the aberrant functional effects of these alterations across ovarian cancer patients from The Cancer Genome Atlas (TCGA) data.

Keywords: Copy number variations, Cancer genomics, Bi–clustering

1 Introduction

Currently, cancer is one of the best characterized diseases from a molecular point of view and, indeed, the wealth of cancer–related data created the opportunity to unravel its development mechanisms [1]. Commonly, the discovery of oncogenes and tumor suppressors is a preliminary step in estimating their impact on altered pathways that cause the onset of cancer [2]. Although many of these genes have been discovered and studied for many types of cancer — they may be either specific for a particular tumour, or active in different tumour types —, usually only their primary functions are recognized [3].

Our investigation is based on the hypothesis that a more systematic and uniform approach to gene–function association might reveal 'secondary' or 'indirect' functional implications. In summary, all processes mediated by oncogenes and tumor suppressors are generally directly or not directly related to different forms of cancer, while a deep knowledge of their interactions is fundamental to better understand the mechanism of oncogenesis.

The present paper proposes a software pipeline which aims at detecting 'weak' (or 'indirect') functions impacted by Copy Number Variations (CNVs) of cancer related genes. CNVs are structural rearrangements involving DNA segments, of at least 50 bp, that can be present with a variable copy number compared to the reference genome [4]. Pathogenic CNVs are often associated with unfavorable cancer prognosis and can occur as specific focal event or as regional aberrations (> 3Mb in length)[5].

CNVs could directly influence the expression pattern of oncogenes and tumour suppressors but their potential contribution to oncogenesis remains poorly understood and unclear. To learn more about the relationships between CNVs and cancer, we applied our pipeline to the task of detecting the aberrant functional effects of oncogenes and tumor suppressors by CNVs on the High Serous Ovarian Carcinoma (HGSOC) dataset, obtained from The Cancer Genome Atlas (TCGA). More precisely, we integrated copy number with gene expression data of 32 oncogenes and 16 tumor suppressors identified by TCGA, as reported in [3]. We have chosen to study this type of cancer because HGSOC is characterized by genomic instability, including recurrent somatic mutations, promoter methylation events and a high frequency of CNVs. In order to determine co–expression profiles among up/down–regulated genes, a recent bi–clustering method [6] is applied, aimed at selecting significant gene activation patterns that are typical only of a subset of patients. Indeed, a standard requirement in analyzing gene data is that of grouping genes according to their expression with respect to multiple patients, since subsets of genes may be co–regulated and co–expressed only for certain patient populations, while they behave almost independently for others. Finding these local expression patterns can be obtained with a two–way clustering or bi–clustering, and is the key to uncover unknown genetic pathways. Finally, among cancer related genes identified by TCGA, we only report results for the oncogene Derlin–1 (DERL1), in order to clearly explain every single step of our pipeline. DERL1 is a component of the endoplasmic reticulum–associated degradation (ERAD) for misfolded proteins and is located on the focally amplified region 8q24.13 identified in HGSOC by TCGA, as reported in [7]. In fact, although DERL1 overexpression is associated with several tumor types, including colon and bladder cancer, the impact of DERL1 CNVs on ovarian cancer neoplasia is not yet well understood.

The paper is organized as follows. In the next section the data collection procedure and the software pipeline, used to detect functional implications of CNVs on the serous ovarian carcinoma, are described, while Section 3 collects experimental results. Finally, in Section 4, we discuss on the obtained outcomes, draw some conclusions and illustrate future perspectives.

2 Materials and methods

2.1 Data sources

The copy number and mRNA expression data of TCGA are collected from the Broad Institute's GDAC Firehose, using the R package RTCGAtoolbox [8]. The copy number calls were determined by TCGA using the GISTIC algorithm [9] which identifies genomic regions that are significantly amplified or deleted across

a set of tumors. Each significantly aberrant region is examined to determine whether it is a focal event (length of the copy number alterations < 3 Mb) or a broad event (equal to the length of a chromosome arm or of an entire chromosome), or shows both characteristics. GISTIC output includes the copy number status of each gene in each sample and it is achieved through the implement of low and high level thresholds to the gene copy levels. The low-level threshold is typically 0.1 or 0.3, while the high-level threshold is computed on a sample by sample basis and is related on the maximum and minimum median arm-level amplification and deletion copy number found in each samples.
Copy number estimation by GISTIC is explained in table 1

Table 1. Copy number estimation by Gistic2

CNV values	Categorization
-1	Shallow Deletion
-2	Deep Deletion
0	Neutral Copy
+1	Low level Gain
+2	High level Amplification

2.2 Pipeline

The core of the proposed pipeline is composed by four steps:

- **Differential analysis**: for a given p–value threshold and absolute log2 fold change threshold, using standard state–of–the–art tools (edgeR [10]), we determine, among a target set of genes, those that are over/under–expressed in the case/control populations;

- **Pathway enrichment analysis**: it is applied to Differentially Expressed Genes (DEGs) using the package ReactomePA [11];

- **GH–EXIN**: In order to further determine co–expressed genes, starting from previously selected up/down–regulated genes, we apply a bi–clustering method [6], aimed at selecting both patients and genes, forming highly homogeneous bi–clusters;

- **GO Enrichment**: for each bi–cluster reaching a user–defined quality threshold and a minimum number of patients, we collect the GO annotations — namely, Biological Process (BP) and Molecular Function (MF) — for all the included genes. A GO category, whose statistical enrichment is below a user–selected p–value threshold, is assumed to be significantly associated with such genes.

2.3 Data differential analysis

Given a pool of patients and a source gene, of which we wish to detect the functional implications of CNVs, we split the patients into two sub–populations: the subpopulation with CNV values = 2, collecting those patients for whom the gene hosts copy number amplifications, and the complementary pool of patients, with CNV values = 0. These two sub–populations form the case/control split under examination. In order to understand the impact and the role of CNVs of DERL1 in ovarian cancer progression we chose to select as a test only the samples with high level of amplifications. Since public repositories of cancer molecular data are often composed of data collected under diverse technical conditions, in order to reduce false positives due to such variability, we performed a batch correction of gene expression levels, using the approach proposed in [4] (also applied in [3]). The differential expression analysis, obtained with the edgeR package, [12] allowed us to identify, among 19990 genes included in our dataset, those genes that are statistically differentially expressed with a p–value <0.01 at an FDR (False Discovery Rate) of 5%.

2.4 The GH–EXIN Neural Network Bi–clustering

In order to further analyze which groups of genes may be affected by CNVs — and based on the assumption that human copy number variations are strictly related to gene expression [13] —, a bi–clustering technique has been applied. Standard clustering algorithms group genes according to their expression on the overall dataset. Bi–clustering algorithms, instead, are capable to find interesting gene activation patterns that are typical only of a subset of patients. Bi–clustering is implemented here in the form of a two–way clustering, in which the clustering operation is alternated in the two dimensions, related to genes and patients, respectively. Only those genes capable to determine cohesive bi–clusters (also with respect to the patient space) must be retained. To this aim, the GH–EXIN neural network clustering has been chosen, a divisive hierarchical clustering algorithm [6, 14] — specifically developed for bi–clustering applications — that builds a tree in an incremental and self–organized way. Clustering algorithms, usually, optimize classic indexes, as the Standard to Noise Ratio (SNR) [2]. Instead, to control the quality of a bi–cluster, GH–EXIN, minimizes the H_{cc} index (introduced in [15]), defined as:

$$H_{cc} = \frac{\sum_i^{N_r} \sum_j^{N_c} r_{ij}^2}{N_r N_c}$$

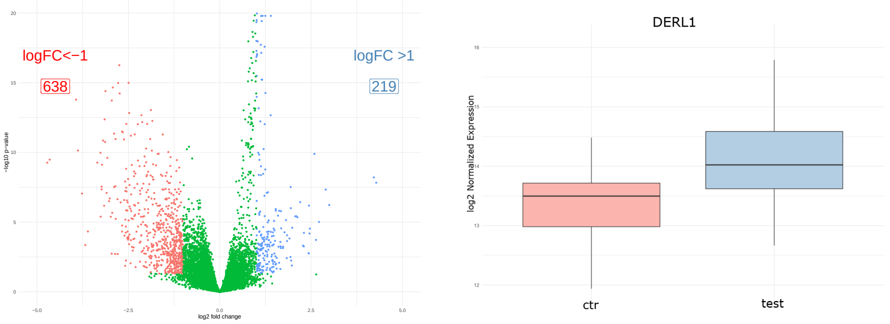
where N_r represents the total number of rows (or genes), N_c represents the total number of columns (or patients), and $r_{ij} = a_{ij} - \sum_k^C a_{ik}/C - \sum_h^R a_{hj}/R + (\sum_i^R \sum_j^C a_{ij})/(CR)$. Matrix A collects the gene expression level for each gene and with respect to each patient, so that a_{ij} is the expression level of gene i in patient j. R and C are the number of rows and columns of the particular bi–cluster, respectively. H_{cc} becomes lower as the bi–cluster tends to be cohesive.

We retain only those groups of genes for which the majority of corresponding H_{cc} indices are below a user–defined threshold.

3 Results

3.1 Identification of DEGs

A total of 3585 genes are identified to be significantly differently expressed between test and control groups, at 5% FDR. Among these DEGs, 219 are up–regulated and 638 are down–regulated, based on a p–value < 0.01 and $|\log 2\ \mathrm{FC}| \geq 1$ as the cut off criteria. Furthermore, our analysis revelead that DERL1 is over–expressed in samples characterized by CNAs (logFC =1.116725, p–value = 2.758823e-46, p–adjust = 4.802008e-42, see Fig. 1).



(a) Volcano plot of differentially expressed genes. Data are plotted as log2 fold change versus -log10 of the adjusted p–value. Blue dots represent up–regulated genes with logFC $\geq$1 and FRD <0.05; red dots represent genes whose expression is down–regulated, with logFC < -1 and FDR 0.05.

(b) Box plots of differential log2–transformed DERL1 count.

Fig. 1. Differential expression analysis.

3.2 Reactome pathway analysis

We further investigated the functional implication of up–regulated and down–regulated genes, with $|\log 2\ \mathrm{FC}| \geq 1.0$, by Reactome pathway analysis. Table 2 shows the top 10 enriched biological pathways with FDR below 0.05. Fig. 2 illustrates relationships between significant genes and the mainly enriched biological pathways — glucuronidation and assembly collagen fibrils and other multimeric structures pathways — as a network.

Table 2. Reactome pathway enrichment analysis.

Pathway	Description	p.value	p.adjust
R-HSA-156588	Glucuronidation	7.97e-13	5.45e-10
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	1.89e-12	6.45e-10
R-HSA-1474244	Extracellular matrix organization	7.06e-12	1.61e-09
R-HSA-1442490	Collagen degradation	2.69e-11	3.93e-09
R-HSA-1474290	Collagen formation	2.87e-11	3.93e-09
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	3.73e-11	4.26e-09
R-HSA-1474228	Degradation of the extracellular matrix	5.81e-11	5.43e-09
R-HSA-8948216	Collagen chain trimerization	6.36e-11	5.43e-09
R-HSA-211859	Biological oxidations	7.52e-09	5.71e-07
R-HSA-156580	Phase II - Conjugation of compounds	8.39e-09	5.74e-07

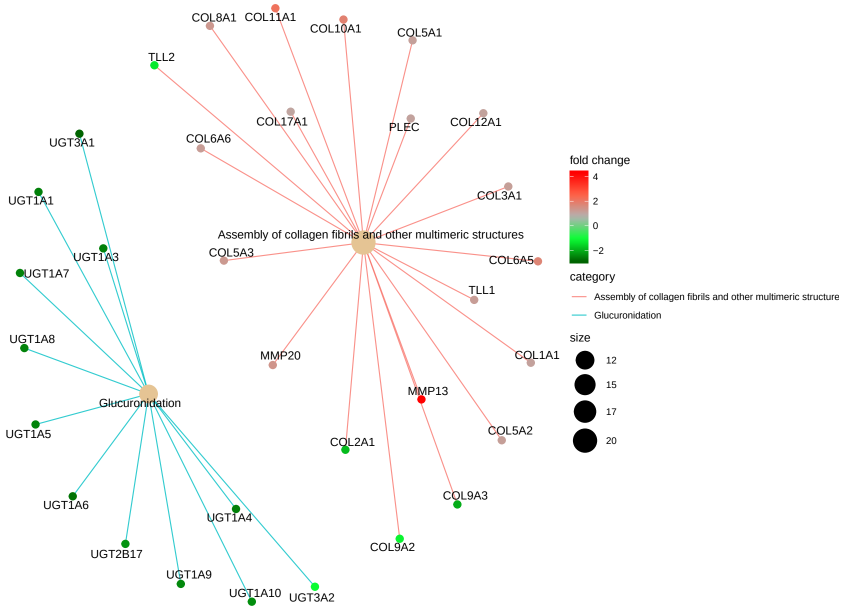


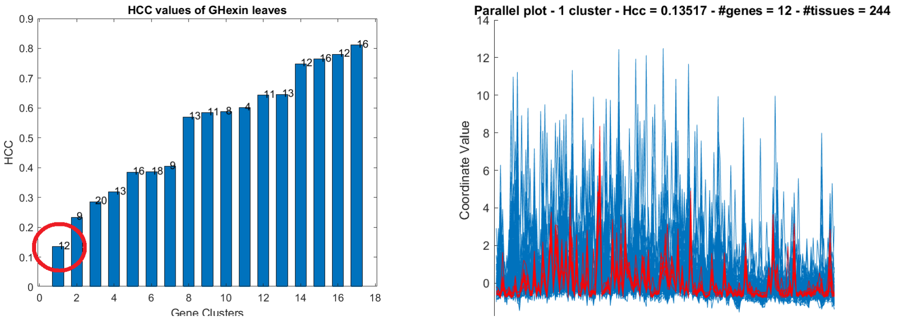
Fig. 2. Pathway gene–network of top 2 enriched Reactome terms from up–regulated and down–regulated genes.

3.3 Bi–clustering results

Subsequently, we have analyzed up–regulated genes with the GH–EXIN bi–clustering algorithm. As a result, a group of 12 genes has been selected. The name of each gene together with the related annotation is reported in Table 3. This group of genes has the lowest H_{cc} value (0.135) among all the groups, as shown in the bar–plot of Fig. 3(a). The expression levels of these genes are also strongly correlated, as shown in the parallel coordinate plot of Fig. 3(b). In this case, the x–axis corresponds to gene expression levels, while the y–axis represents the patients.

Table 3. DERL1 cluster of up–regulated genes.

Gene	annotation
FAP	Prolyl endopeptidase FAP
ADAM12	Disintegrin and metalloproteinase
COL8A1	Collagen alpha-1(III) chain
COL1A1	Collagen alpha-1(I) chain
SLC35D3	Solute carrier family 35 member D3
COL3A1	Collagen alpha-1(III) chain
COL5A1	Collagen alpha-1(V) chain
INHBA	Inhibin beta A chain
FN1	Fibronectin type III
ALPK2	Alpha-protein kinase
THBS2	Thrombospondin-2
COL5A2	Collagen alpha-2(V) chain



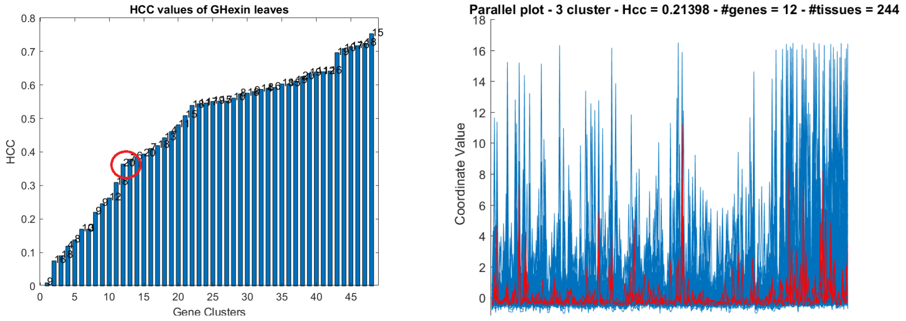
(a) H_{cc} values of all gene clusters produced by GH–EXIN. In the red circle, the group of 12 selected genes.

(b) Parallel coordinates of the expression levels of all up–regulated genes in the selected clusters. In red, genes selected by the GH–EXIN algorithm.

Fig. 3. Bi–clustering analysis on the up–regulated genes.

Table 4. DERL1 cluster of down–regulated genes.

Gene	Annotation
PCSK9	Proprotein convertase subtilisin/kexin type 9
SLC47A1	Multidrug and toxin extrusion protein 1;
NUDT11	Diphosphoinositol polyphosphate phosphohydrolase 3-beta
ADAMTS19	ADAM metallopeptidase with thrombospondin type 1 motif 19
BAMBI	BMP and activin membrane-bound inhibitor homolog
MSX1	Homeobox protein MSX-1
FGF9	Fibroblast growth factor 9
FGF19	Fibroblast growth factor 19
UGT3A1	UDP-glucuronosyltransferase 3A1
SMTNL2	Smoothelin-like protein
CLDN2	cell-adhesion activity
NOTUM	Palmitoleoyl-protein carboxylesterase
TMEM59L	Transmembrane protein 59-like;
IHH	Indian hedgehog protein
EN2	Homeobox protein engrailed-2
POMC	Pro-opiomelanocortin; Endogenous opiate
SHISA6	Protein shisa-6 homolog
DKK4	Dickkopf-related protein; Antagonizes Wnt
RIMKLA	N-acetylaspartylglutamate synthase A



(a) H_{cc} values of all gene clusters produced by GH–EXIN. In the red circle, the group of 20 selected genes.

(b) Parallel coordinates of the expression levels of all down–regulated genes in the selected clusters.

Fig. 4. Bi–clustering analysis on the down–regulated genes.

Following the same approach, we have also analyzed down–regulated genes. Results of this analysis are reported in Figs. 4(a)–4(b). As it is shown in Fig. 4(a), the group of gene we choose in this case is not the one at minimum H_{cc}. Sure enough, the group of genes still has a low H_{cc} value, but there were groups with lower values. Anyway, as shown in Section 3.4, the GO enrichment analysis revealed more interesting relationships within this group with regards to the others. The selected genes are summarized in Table 4.

3.4 GO enrichment Cluster

Biological significance of selected clusters are explored by the GO term enrichment analysis, including molecular function (MF) and biological process (BP), by using a p–value < 0.01 and an FDR below 0.05, as the cut off criteria. Up–regulated genes included in the DERL1 cluster are enriched in 13 terms in the category MF and in 25 in the category BP. In the molecular function group, up–regulated genes are mainly enriched in extracellular matrix structural constituent (p–value= 8.84E-12, p–adjust= 4.33E-10) and in platelet–derived growth factor binding (p–value=1.45E-10,p–adjust= 3.56E-09). In the biological process group, up–regulated genes are mainly enriched in extracellular matrix organization (p–value = 4.66E-11, p–adjust= 2.22E-08) and in endodermal cell differentiation (p–value = 2.02E-10, p–adjust= 3.21E-08).

Based on the same approach, down–regulated genes included in the DERL1 cluster are enriched in 3 terms in the category MF and in 4 in the category BP. In the molecular function group, down–regulated genes are mainly enriched in receptor regulator activity (p–value= 7.332415e-05, p–adjust= 0.004472773) and in fibroblast growth factor receptor binding (p–value= 1.950151e-04, p–adjust= 0.005947961). In the biological process group, down–regulated genes are mainly enriched in regulation of canonical Wnt signaling pathway (p–value= 1.126880e-05, p–adjust= 0.005490312) and in mesenchymal cell proliferation (p–value = 1.854439e-05, p–adjust= 0.005490312)

4 Discussion and conclusions

Ovarian cancer is the fifth leading cause of cancer death among women aged 35 to 74. In particular, high–grade serous ovarian cancer (HGSOC) is the most aggressive type, showing the lowest survival rate and low response to standard therapy. The genomic analyses carried out by the TCGA research network revealed that HGS–OVCa is characterized by 113 significant focal DNA copy number aberrations. Interestingly, the SNP copy–number analysis performed by TCGA identified an amplification region at 8q24.3, that includes the oncogene DERL1, a component of ERAD pathway. From these evidence, in this study, we applied our pipeline to shed light on the indirect effect of copy number amplifications of the oncogene DERL1 in HGSOC development. We found that samples characterized by the CNAs of DERL1 showed 3585 differently expressed genes in respect to the control group, including DERL1. According to Reactome enrichment analysis, the DEGs are mainly enriched in glucuronidation and assembly collagen fibrils and other multimeric structures pathways, fundamental processes for the cancer development. Deregulation of glucuronidation process, through the down–regulation of all components of UDP–glucuronosyl transferase (UGTs) protein family, could lead to the accumulation of cellular compounds and improve the cellular fitness, allowing the tumour to growth and cells to proliferate under stressful conditions. Within the pathway of ”assembly collagen fibrils and other multimeric structures” there are both down–regulated and up–regulated genes which can contribute to tumour malignancy and increase cancer cells proliferation. Indeed, several studies have demonstrated that, in pathological conditions, the down–regulation of Collagen alpha–1(II)chain(COL2A1) gene, responsible for the production of the alpha1(II) chain of type II collagen, could promote cellular proliferation and migration [16]. At the same time, the over–expression of Matrix metallopeptidase 13 (MMP13), implicated in the breakdown of extracellular matrix, could induce cancer metastasis through known mechanisms as tumour cell invasion [17]. In addition, the bi–clustering approach allowed us to identify the co–expression profiles among significantly genes. These genes have important roles in biological processes which, if deregulated, could contribute to cancer progression. In particular, the evidenced clusters are involved in extracellular matrix organization and in regulation of Wnt signal transduction, key mechanisms in cancer biology. These results suggest us that the copy number amplifications of oncogene DERL1 and its over–expression between the two conditions could influence the expression of other genes involved in fundamental pathways for ovarian cancer progression.

Different genes other than DERL1 are included in the focal amplification region 8q24.3 and they could also influence our analyses, making difficult to understand the real impact of DERL1 aberration in ovarian cancer. Further analysis will be necessary to clarify the contribute of DERL1 amplification in this type of cancer. Concluding, a deeper bioinformatic and gene co–expression network analysis is required to confirm the impact of CNVs of DERL1 and to investigate the function of clusters associated with the development of HGSOC.

References

1. Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299, 2014.
2. Louise B Thingholm, Lars Andersen, Enes Makalic, Melissa C Southey, Mads Thomassen, and Lise Lotte Hansen. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: addressing the challenges. *Frontiers in genetics*, 7:2, 2016.
3. Qingyang Zhang, Joanna E Burdette, and Ji-Ping Wang. Integrative network analysis of tcga data for ovarian cancer. *BMC systems biology*, 8(1):1338, 2014.
4. Lorenzo Tattini, Romina D'Aurizio, and Alberto Magi. Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, 3:92, 2015.
5. Laia Rodriguez-Revenga, Montserrat Mila, Carla Rosenberg, Allen Lamb, and Charles Lee. Structural variation in the human genome: the impact of copy number variants on clinical diagnosis. *Genetics in Medicine*, 9(9):600, 2007.
6. B. Pietro, C. Gabriele, E. Piccolo, C. Giansalvo, C. Maurizio, and A. Bertotti. Neural biclustering in gene expression analysis. In *2017 International Conference*

on Computational Science and Computational Intelligence (CSCI), pages 1238–1243, Dec 2017.

7. Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.
8. Mehmet Kemal Samur. Rtcgatoolbox: a new tool for exporting tcga firehose data. *PloS one*, 9(9):e106397, 2014.
9. Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, 2011.
10. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
11. Guangchuang Yu and Qing-Yu He. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12(2):477–479, 2016.
12. Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.
13. Stranger B. E. Gamazon, E. R. The impact of human copy number variation on gene expression. *Briefings in functional genomics*, 14(5):352—357, 2015.
14. Giansalvo Cirrincione, Gabriele Ciravegna, Pietro Barbiero, Vincenzo Randazzo, and Eros Pasero. The gh-exin neural network for hierarchical clustering. *Neural Networks*, 121:57 – 73, 2020.
15. Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.
16. Yihang Qi, Xiangyu Wang, Xiangyi Kong, Jie Zhai, Yi Fang, Xiaoxiang Guan, and Jing Wang. Expression signatures and roles of micrornas in inflammatory breast cancer. *Cancer Cell International*, 19(1):23, Jan 2019.
17. Yasusei Kudo, Shinji Iizuka, Maki Yoshida, Takaaki Tsunematsu, Tomoyuki Kondo, Ajiravudh Subarnbhesaj, Elsayed M Deraz, Samadarani BSM Siriwardena, Hidetoshi Tahara, Naozumi Ishimaru, et al. Matrix metalloproteinase-13 (mmp-13) directly and indirectly promotes tumor angiogenesis. *Journal of Biological Chemistry*, 287(46):38716–38728, 2012.