Large scale statistical analysis of GEO datasets

Bernard Ycart, Konstantina Charmpi, Sophie Rousseaux, Jean-Jacques Fournié

Large scale statistical analysis of GEO datasets

Bernard Ycart*[1,2,3] , Konstantina Charmpi[1,2,3], Sophie Rousseaux[1,4], Jean-Jacques Fournié[3,5,6,7]

[1] Université Grenoble Alpes, France
[2] Laboratoire Jean Kuntzmann, CNRS UMR5224, Grenoble, France
[3] Laboratoire d'Excellence TOUCAN, France
[4] INSERM, UMR823, Institut Albert Bonniot, Grenoble, France
[5] INSERM UMR1037-Cancer Research Center of Toulouse, Toulouse, France
[6] Université Toulouse III Paul-Sabatier, Toulouse, France
[7] ERL 5294 CNRS, Toulouse, France

Email: Bernard Ycart*- Bernard.Ycart@imag.fr; Konstantina Charmpi - Konstantina.Charmpi@imag.fr; Sophie Rousseaux - Sophie.Rousseaux@ujf-grenoble.fr; Jean-Jacques Fournié - Jean-Jacques.Fournie@inserm.fr;

*Corresponding author

Abstract

The problem addressed here is that of simultaneous treatment of several gene expression datasets, possibly collected under different experimental conditions and/or platforms. Using robust statistics, a large scale statistical analysis has been conducted over 20 datasets downloaded from the Gene Expression Omnibus repository. The differences between datasets are compared to the variability inside a given dataset. Evidence that meaningful biological information can be extracted by merging different sources is provided.

Background

Many genomewide expression datasets have been published during the past ten years. Repositories, such as the Gene Expression Omnibus (GEO) database [1], have made available an impressive wealth of data. Using them as a whole, instead of restricting statistical studies to one particular dataset, is tantalizing. Two recently published R/Bioconductor packages [2, 3] provide various tools for merging datasets coming from different studies. However, a serious doubt has been cast by Haibe-Kains et al. [4], after comparing two large scale pharmacogenomic studies: whereas both studies had a good overall correlation, important discordances could be observed. Thus, the following crucial question remains to be answered: is it statistically legitimate to merge datasets coming from different studies? An attempt at answering this question is reported here.

Merging different datasets, requires prior checking that the information they contain is compatible, and hence that detected differences between gene expressions under different conditions are not artifacts, due to experimental or data processing methods. An obvious obstacle to simultaneous treatment is that expression data collected under different experimental conditions and/or platforms usually have incompatible

distributions, which differ sometimes by several orders of magnitude [5,6]. A solution is provided by robust (or distribution-free) statistics [7,8]. Robust methods amount to replacing actual values by ranks, or equivalently by empirical distribution functions or van der Waerden's normal scores [7, p. 309]. This idea has already been applied to expression data in several papers, including [9–11]. However, to the best of our knowledge, a large scale analysis assessing the reproducibility of information from one dataset to another, is still missing. We have conducted such an analysis over 20 GEO datasets, totalling 17 745 genomewide expression samples.

For the data treatments presented here, the statistical language R [12] has been used. Our set of functions, together with a manual, has been made available online as supplementary information. Throughout the article, we consider *data matrices* (also called assay data in [13]) as containing expression data relative to a set of genes. Each row corresponds to a different gene symbol, or *feature*, each column to a different data vector or *sample* (see Table 2 in [1]). Such a matrix is deduced from raw datasets, available on the GEO repository, though standard treatments: annotation and reduction [14]. Several R packages [15, 16] that perform these operations and output data matrices such as considered here, are available. We have encoded our own functions. We have chosen a data structure in which each data matrix is paired with its *information matrix*. The columns of the information matrix are labelled by the same numbers as the paired data matrix. Its rows contain the different information fields of the data. Our focus here is on overexpression or underexpression of genes, in different tissues or cancer types.

Our objective was twofold. On the one hand, we wanted to check whether the information on genes, contained in different data matrices, was compatible, and to which extent. This was done on a set of 20 different matrices. Various statistical treatments were performed. The first one consisted in computing correlations between median columns of the matrices. Vectors of pairwise correlations between rows were also compared. Then multivariate analysis over assays of gene symbols was applied: Wilcoxon and Kruskal-Wallis tests, factor and principal component analysis (PCA). The results were compared to those obtained by sorting a single matrix according to different keywords. All comparisons showed not perfect, but highly significant correlations. However, it was also found that in all cases, a sizeable proportion of symbols were good discriminators of the different matrices. But this was also found to hold between two submatrices inside a given dataset. Therefore, it cannot be regarded as an obstacle to merging different datasets. On the other hand, we wanted to know whether biological information could be consistently retrieved from matrices collated from different sources. Two merged sets of matrices were made. The first one came from general cancer cell datasets, from which samples of breast and lung tumors were extracted. The second one was made of blood RNA datasets, coming from healthy individuals, or from leukemias. In both cases, evidence that already known biological information could be extracted from merged matrices was found.

Methods

The datasets available from the GEO repository [1], collate sets of expression vectors, or samples. Several R/Bioconductor packages can be used to download and format the data [15, 16]. We have chosen to encode in R our own functions. Our R script has been made available online, together with a user manual. Our formatting choices are described below.

In a GEO dataset two types of information are available for each sample. The first type consists of numeric values corresponding to a set of probes. The second type are character-type informations on the

experimental setting. We have chosen to separate the two types into data matrices and information matrices. In the data matrix, probes are associated to gene symbols with the use of different Bioconductor annotation packages according to the platform [17–20]. After annotation, some symbols are duplicated. Several methods can be used to eliminate duplicates. We have chosen to keep the row with the largest interquartile range, as in [14], because we believe that this is the most statistically coherent choice. After annotation and reduction, the data matrix, with gene symbols as row names, and series numbers as column names, is saved as a single R object for future use. The information matrix has the same column names as the corresponding data matrix. Its rows correspond to the different fields.

Our merging function reduces data matrices to common row names. For information matrices, different sets of data usually have different information fields. This was taken into account when merging two information matrices, by indexing the rows of the merged matrix by the union of row names in the initial information matrices.

Two R/Bioconductor packages have recently been issued for merging GEO datasets [2, 3]. In [2], quantile discretization, normal discretization normalization, gene quantile normalization, median rank scores, quantile normalization (QN) are proposed. In [2, 3], the Batch Mean-Centering method, Distance-Weighted Discrimination, Z-score standardization, and the Cross-Platform Normalization method are proposed. An Empirical Bayes (EB) method is available in both packages. For the results reported here, only classical methods were used, and we consider them as sufficient to establish our main points, our focus being on overexpression or underexpression of genes, in different tissues or cancer types.

As in [9–11], we have made the choice to use robust statistics [7, 8]. This implies changing the columns of a data matrix into distribution free values. The usually proposed transformation replaces the i-th value x_i by its rank R_i if x_i is the R_i-th smallest value in the column. However, ranks range between 1 and the number of rows. The problem is that different matrices may have different numbers of rows (gene symbols). In order to get a unique range of values for all matrices, it seems preferable to use a scale free score. The simplest such score is the Empirical Cumulative Distribution Function (ECDF): its value at x_i is R_i/n, if n is the number of rows. Graphical displays look more familiar if another score is used: the van der Waerden's normal score [7, p. 309]. It consist of replacing x_i by $\phi(R_i/(n+1))$, where ϕ is the quantile function of the standard normal distribution. With the ECDF, the distribution of each column becomes uniform on the interval $[0, 1]$, whereas with the normal score it becomes standard normal. Results reported above have been obtained with the normal score, but they are not essentially different if the ECDF is used instead.

In statistical inference, the choice of robust statistics must be made coherent. This is the reason why we have replaced the usual normal-sample techniques by their robust equivalent, and used medians instead of means, Spearman's correlation instead of Pearson's [7, p. 422-431], Wilcoxon (or Mann-Whitney) location test instead of Student's t-test [7, p. 268-278], Kruskal-Wallis test instead of one-way analysis of variance [7, p. 363-372]. When comparing several matrices to detect location diffferences, the Kruskal-Wallis test was run over all common rows. When differentiating overexpression from underexpression, a one-sided Wilcoxon test was run. The same test being used for a large number of features, a False Detection Rate (FDR) correction of p-values by the Benjamini-Yekutieli method [21] was systematically applied. Features were ranked from most to least significant, either by sorting p-values in increasing order, or by sorting the values of the test statistic instead. We considered as significant, any feature with a (FDR-corrected) p-value smaller than 5%. Once a set of (significant) features had been selected, the corresponding rows were concatenated into single

vectors. These vectors were taken as variables, and the samples as individuals, for a PCA. Figures 3 to 5 were obtained by projecting the samples as points onto the first principal plane, and differentiating their initial data matrices by colors. Precise R commands can be found in the user manual made available online.

Results

The 20 datasets that were downloaded from the GEO repository are detailed in Table 1. They were selected on a criterion of size (number of samples: 500 or more). The 20 matrices together amount to 17 745 samples. To each study, a three-letter acronym was attached; these acronyms will be used in what follows.

acronym	reference	series number	platform number	symbols (rows)	samples (columns)
EPO	[22]	2109	570	20 184	2158
PMM	[23]	2658	570	20 184	559
AML	[24]	6891	570	20 184	537
HBI	[25]	7307	570	20 184	677
MIL	[26]	13159	570	20 184	2 096
MDS	[27]	15061	570	20 184	870
PLE	[28]	20142	6947	19 626	1 240
MMD	[29]	24080	570	20 184	559
DLB	[30]	31312	570	20 184	498
PRS	[31]	33828	10558	20 768	881
CCL	[32]	36133	15308	18 722	917
BEC	[33]	36192	6947	19 628	911
WBS	[34]	36382	6947	19 628	991
GSC	[35]	36809	570	18 260	812
MBI	[36]	37069	570	18 260	590
CCC	[37]	39582	570	20 184	566
PVA	[38]	48152	6947	19 628	705
HPS	[39]	48348	6947	19 628	734
XMD	[40]	48433	570	20 184	823
HAV	[41]	48762	6947	19 628	621

Table 1: Twenty GEO series have been chosen, coming from four different platforms. To each of them a three letters acronym was associated. The table gives the acronym, a recent reference, the GEO series number, the platform number. For the data matrix (or assayData), the number of symbols after annotation and reduction, and the number of columns (samples) are given. All 20 data matrices had 15 562 gene symbols in common.

In the results reported here, each data matrix has been transformed by replacing its column values, by the corresponding van der Waerden normal scores [7, p. 309]. Similar results were obtained when replacing column values by their empirical distribution function (see methods section).

The first treatment that was applied consisted in computing, for each dataset, the median of all rows, reduced to the 15 562 common gene symbols. This gave 20 vectors of length 15 562, the correlation matrix of which is given in Table 2. A positive (negative) correlation between vectors of size 15 562 is significant at threshold 5% if it is larger than 0.013 (smaller than -0.013); thus all correlations of Table 2 can be regarded as significant.

Figure 1 shows a factor analysis of the 20 variables. Fifteen of them can be clustered into four groups.

	EPO	PMM	AML	HBI	MIL	MDS	PLE	MMD	DLB	PRS
EPO	1.00	0.80	0.71	0.92	0.63	0.82	0.55	0.48	0.59	0.18
PMM	0.80	1.00	0.75	0.79	0.63	0.81	0.56	0.63	0.55	0.19
AML	0.71	0.75	1.00	0.68	0.76	0.85	0.61	0.58	0.59	0.18
HBI	0.92	0.79	0.68	1.00	0.58	0.76	0.53	0.46	0.53	0.18
MIL	0.63	0.63	0.76	0.58	1.00	0.78	0.58	0.67	0.61	0.14
MDS	0.82	0.81	0.85	0.76	0.78	1.00	0.69	0.52	0.60	0.18
PLE	0.55	0.56	0.61	0.53	0.58	0.69	1.00	0.40	0.45	0.22
MMD	0.48	0.63	0.58	0.46	0.67	0.52	0.40	1.00	0.50	0.11
DLB	0.59	0.55	0.59	0.53	0.61	0.60	0.45	0.50	1.00	0.12
PRS	0.18	0.19	0.18	0.18	0.14	0.18	0.22	0.11	0.12	1.00
CCL	0.81	0.74	0.69	0.75	0.65	0.77	0.55	0.54	0.56	0.20
BEC	0.56	0.49	0.43	0.65	0.38	0.48	0.63	0.32	0.35	0.21
WBS	0.57	0.58	0.62	0.54	0.58	0.70	0.94	0.40	0.46	0.23
GSC	0.59	0.61	0.69	0.57	0.63	0.74	0.66	0.46	0.49	0.15
MBI	0.60	0.63	0.70	0.58	0.65	0.75	0.64	0.48	0.50	0.15
CCC	0.77	0.62	0.64	0.67	0.59	0.67	0.50	0.49	0.53	0.15
PVA	−0.09	−0.10	−0.13	−0.09	−0.14	−0.16	−0.30	−0.09	−0.09	0.16
HPS	0.62	0.62	0.66	0.58	0.62	0.74	0.93	0.43	0.49	0.24
XMD	0.90	0.81	0.72	0.84	0.65	0.83	0.56	0.51	0.56	0.19
HAV	0.60	0.60	0.64	0.56	0.61	0.73	0.89	0.42	0.48	0.21

	CCL	BEC	WBS	GSC	MBI	CCC	PVA	HPS	XMD	HAV
EPO	0.81	0.56	0.57	0.59	0.60	0.77	−0.09	0.62	0.90	0.60
PMM	0.74	0.49	0.58	0.61	0.63	0.62	−0.10	0.62	0.81	0.60
AML	0.69	0.43	0.62	0.69	0.70	0.64	−0.13	0.66	0.72	0.64
HBI	0.75	0.65	0.54	0.57	0.58	0.67	−0.09	0.58	0.84	0.56
MIL	0.65	0.38	0.58	0.63	0.65	0.59	−0.14	0.62	0.65	0.61
MDS	0.77	0.48	0.70	0.74	0.75	0.67	−0.16	0.74	0.83	0.73
PLE	0.55	0.63	0.94	0.66	0.64	0.50	−0.30	0.93	0.56	0.89
MMD	0.54	0.32	0.40	0.46	0.48	0.49	−0.09	0.43	0.51	0.42
DLB	0.56	0.35	0.46	0.49	0.50	0.53	−0.09	0.49	0.56	0.48
PRS	0.20	0.21	0.23	0.15	0.15	0.15	0.16	0.24	0.19	0.21
CCL	1.00	0.56	0.56	0.56	0.59	0.74	−0.08	0.61	0.92	0.60
BEC	0.56	1.00	0.64	0.37	0.37	0.45	−0.13	0.65	0.57	0.59
WBS	0.56	0.64	1.00	0.64	0.63	0.49	−0.29	0.95	0.57	0.88
GSC	0.56	0.37	0.64	1.00	0.94	0.61	−0.19	0.66	0.57	0.64
MBI	0.59	0.37	0.63	0.94	1.00	0.62	−0.17	0.66	0.59	0.64
CCC	0.74	0.45	0.49	0.61	0.62	1.00	−0.09	0.53	0.75	0.51
PVA	−0.08	−0.13	−0.29	−0.19	−0.17	−0.09	1.00	−0.27	−0.08	−0.26
HPS	0.61	0.65	0.95	0.66	0.66	0.53	−0.27	1.00	0.62	0.90
XMD	0.92	0.57	0.57	0.57	0.59	0.75	−0.08	0.62	1.00	0.60
HAV	0.60	0.59	0.88	0.64	0.64	0.51	−0.26	0.90	0.60	1.00

Table 2: For each of the 20 data matrices of Table 1, the median column value of each gene symbol was computed. This gave 20 vectors with length 15 562 (number of common symbols). The table gives pairwise correlations between the 20 vectors.

- PMM, EPO, XMD, HBI, CCL, CCC. Among these six datasets, four are generalist studies involving different tissues and conditions (EPO, HBI, XMD, CCL); CCC concerns colon cancer, and PMM multiple myelomas. Observe that CCL, which was obtained under a platform different from the five others, has excellent correlations with them (between 0.74 and 0.92).
- WBS, PLE, HPS, HAV. All four correspond to blood RNA samples from healthy patients.
- MIL, AML, MDS. All three correspond to leukemias.
- GSC, MBI. These two matrices correspond to similar tissues (blood samples), and similar conditions (critical injuries and burn injuries). Moreover, they were produced on the same platform, by the same

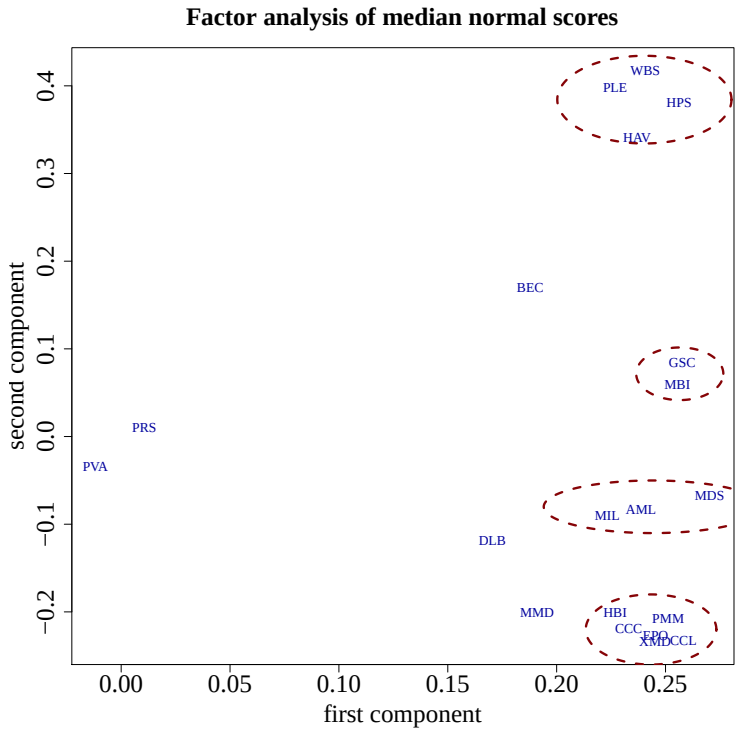

Figure 1: Factor analysis of median columns for 20 datasets. The 20 variables are projected onto the first principal plane of the PCA. Four clusters are identified.

organization. Their excellent correlation (0.94) is not a surprise.

Three datasets, BEC, DLB, MMD have relatively good correlations with those of the above four groups (around 0.5), but no particular links with those groups, nor between themselves. The relative surprise comes from the weak correlations of PRS, and the negative correlations of PVA. Both come from blood RNA samples, and they could have been expected to be close to the WBS, PLE, HPS, HAV group. That PRS and PVA are far from any other matrix, can be explained by their inner heterogeneity. It is illustrated for PVA on Figure 2, where the values over features ALPP and CA4 are represented: samples separate into 4 clusters, according to over- or underexpression of the two genes. As an example, if PVA is split into samples for which the value of ALPP is positive (overexpression), or negative (underexpression), and the row medians are calculated over the two submatrices as before, a correlation of -0.69 is found: thus one half of PVA has a strong negative correlation with the other half. Similar results are obtained for many other features. We considered that the heterogeneity of PVA and PRS did not qualify them for merging.

For each matrix, we also computed all possible pairwise row correlations: 20 vectors of more than 121 millions of pair-correlations were obtained: this is the technique used to evaluate genes for cross-platform consistency of expression patterns in [42]. As expected, the correlation matrix had smaller values than that of Table 2. For instance, the correlation of CCL with XMD was 0.53 instead of 0.92, but still highly significant because of the large number of values.

Correlations between column medians or pair-correlations, is too crude a criterion to judge the homo-

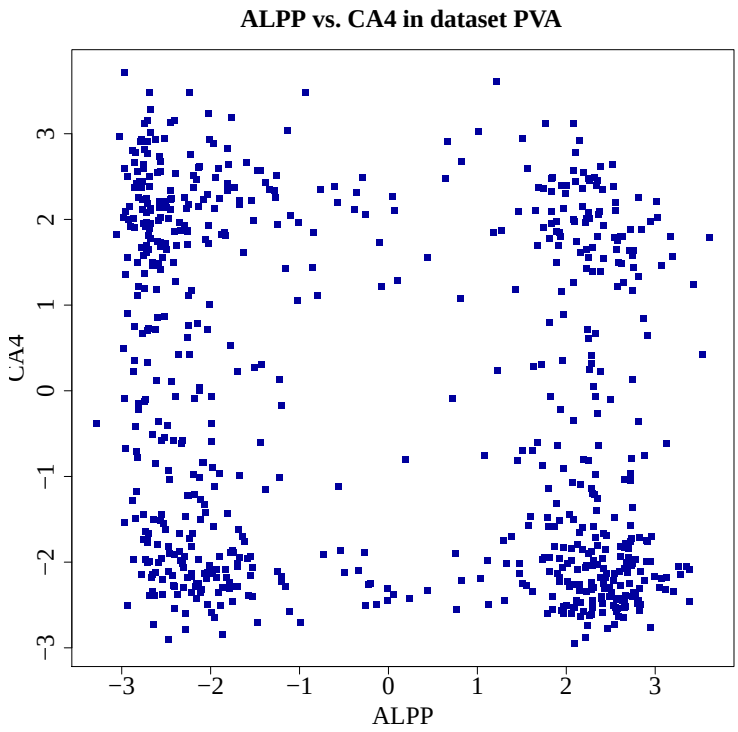

Figure 2: Values of PVA on CA4 versus ALPP. Samples separate into 4 clusters, according to over- or underexpression of the two genes.

geneity of two datasets. As an example, GSC and MBI have an excellent median correlation of 0.94, and several good reasons to be similar. Yet, when each feature is tested for significant differences by the Kruskal-Wallis test, 14 800 significant features out of 18 260 are detected (see methods section for details). The same occurred for any pair of datasets: the distributions of rows had significantly different location parameters, for a majority of features. This means that, for a majority of genes, the ranks of their expressions in the first dataset are significantly smaller or larger than in the second.

Since discrepancies appear to be observed between any two datasets, it must be decided whether they are due to actual biological information, or to a statistical artifact, induced by the experimental setting or the platform. For this, we focused on the dataset MIL (GSE13159 [26]), that has 2 096 samples. The samples were sorted into six submatrices, according to six keywords: Healthy (74 samples), ALL (acute lymphoblastic leukemia, 750 samples), AML (acute myeloid leukemia, 542 samples), CLL (chronic lymphocytic leukemia, 448 samples), CML (chronic myelogenous leukemia, 750 samples), MDS (myelodysplastic syndrome, 202 samples). Then the same treatments as before were applied. Firstly the six median columns were computed, and their correlation matrix was obtained (Table 3).

The values are between 0.85 and 0.99, which is in the range of the best correlations of Table 2. As a control, we made a partition of the same matrix into 6 random subsets, with the same numbers of samples as above, and computed the correlation matrix in the same way. On the control random partition, all correlations were above 0.997. This proves that the partition into keywords does contain meaningful differences. Indeed,

	Healthy	ALL	AML	CLL	CML	MDS
Healthy	1.00	0.92	0.96	0.86	0.98	0.99
ALL	0.92	1.00	0.95	0.91	0.90	0.91
AML	0.96	0.95	1.00	0.89	0.96	0.97
CLL	0.86	0.91	0.89	1.00	0.84	0.85
CML	0.98	0.90	0.96	0.84	1.00	0.98
MDS	0.99	0.91	0.97	0.85	0.98	1.00

Table 3: The data matrix MIL was partitioned according to the 6 keywords Healthy, ALL, AML, CLL, CML, MDS. For each of the six submatrices, the median column of each feature was computed. This gave 6 vectors with length 20 184 (number of symbols in MIL). The table gives the correlations of the 6 vectors.

these differences were detected by the Kruskal-Wallis test: out of the 20 184 features, 18 301 were found significant. Twenty-two features had Kruskal-Wallis p-value below 10^{-300}: SOX4, SYNGR2, ERLIN1, FAH, C7orf23, PSMA6, RTN3, UHRF1, ADAM28, BLK, FUCA2, CD79A, ADA, MYL6B, HEBP1, LEF1-AS1, LEF1, AFF3, COL9A2, MICALL2, MPO, PPM1K. A PCA of the corresponding rows of MIL was run, and the samples projected as points onto the first principal plane, differentiating submatrices by colors (Figure 3). The two submatrices ALL (blue points) and CLL (brown points) are clearly separated from the rest.

Differences inside a given dataset can be induced by several factors. Two factors may not induce differences of the same order of magnitude. However, there is no statistical reason why a dataset like MIL should not be used as a whole, and many ways to verify that the observed differences correspond to actual biological information. Here is an example. Stirewalt et al. [43] list a group of 7 genes displaying increased expression in acute myeloid leukemia samples: BIK, CCNA1, FUT4, IL3RA, HOMER3, JAG1, WT1. When a one-sided Wilcoxon test is applied to the submatrix AML versus the rest of MIL, those 7 genes are among the most significant: their p-values range between 6.7×10^{-130} and 4.3×10^{-42}. The most significant, HOMER3, ranks 54-th among the 20 184 features of MIL.

If observed differences between two datasets (like GSC and MBI) are of the same order of magnitude as differences inside a given dataset, such as caused by a significant factor (see figure 3), it can be admitted as statistically legitimate to merge the two datasets. That meaningful information can be obtained from the merging, remains to be proved. In the following experiments, matrices to be merged were selected in the clusters detected by factor analysis (Figure 1).

Our first experiment consisted in extracting samples corresponding to breast and lung tumors, from the three matrices CCL, EPO, and XMD. CCL has 56 samples of breast tumors, and 166 of lung tumors, EPO has 367 and 143, XMD has 32 and 152. Two matrices "Breast" and "Lung" were made by merging the six submatrices three by three, according to tissues. They had 18 466 features in common, by 455 samples for Breast, and 461 for Lung.

The Kruskal-Wallis test was run on the six separated submatrices, then on the two matrices Breast and Lung. The ten most significant symbols were extracted, and a PCA was run as before. The results are displayed on Figure 4. Significant symbols when the 6 matrices are separated (left panel) are different from significant symbols separating Breast and Lung (right panel). On the left panel, it is clear that the information on the dataset (CCL, EPO, or XMD) dominates the separation Breast vs. Lung: samples coming from CCL are on the left, from EPO on the right, from CCL in the middle. But on the right panel,

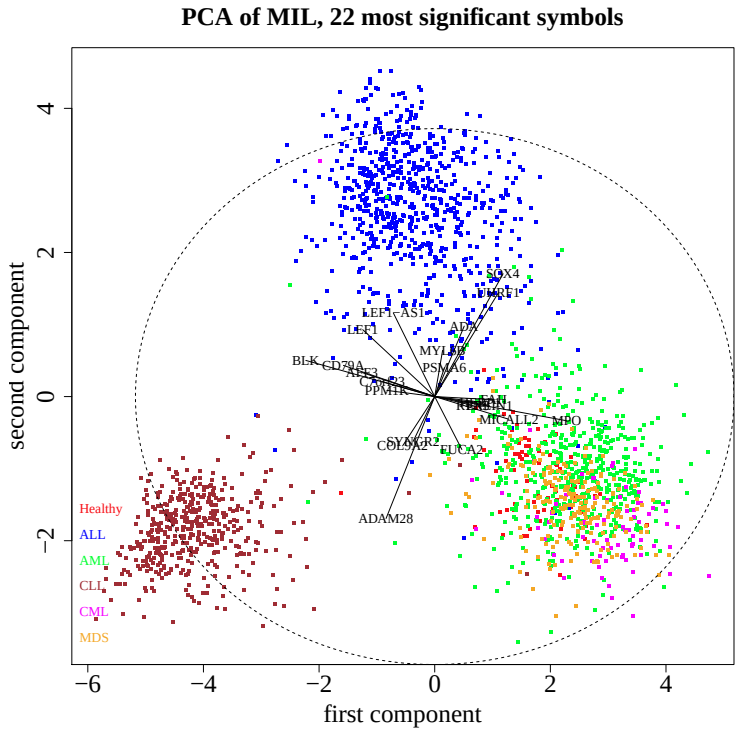

Figure 3: Dataset MIL, partitioned into 6 submatrices according to keywords Healthy, ALL, AML, CLL, CML, MDS. PCA of the 22 symbols with Kruskal-Wallis p-value under 10^{-300}: SOX4, SYNGR2, ERLIN1, FAH, C7orf23, PSMA6, RTN3, UHRF1, ADAM28, BLK, FUCA2, CD79A, ADA, MYL6B, HEBP1, LEF1-AS1, LEF1, AFF3, COL9A2, MICALL2, MPO, PPM1K. Samples are represented by points, with six different colors.

the two types of tumors are also clearly separated. Separators include GATA3 on the right side (Breast), IGF2BP3 on the left side (Lung). Two articles, among others, show the importance of GATA3 for breast cancer [44, 45]. In [46], the link of IGF2BP3 to lung cancer is explicitly stated.

Further information was obtained by running a one-sided Wilcoxon test to detect symbols separating both types of tumor. Then the Molecular Signature database C2 [47] was searched for symbols matching them. Among the 20 genes found most significantly overexpressed in breast tumors by our test, 11 were inside genesets of C2 relative to breast cancers, and outside all genesets relative to lung tumors: EFHD1, IRX5, MUCL1, PRLR, PTGER3, RGL2, TRIL, TRPS1, VAV3, WWP1, ZG16B. Seven of these genes can be found in the G2SBC database [48] and for 10 out of 11, we have found at least one reference relating it to breast cancer. Conversely, among the most significant genes for lung tumor, the following were found in C2 genesets related to lung and not in those related to breast: ALDH3B1, DARS, PRPSAP2, FAM96B, MBIP, LRRC20. The overexpression of ALDH3B1 in lung tumors has been reported in [49]. Santarius et al. [50] gives lists of genes, the overexpression of which is associated to different types of human cancers. The genes detected as significantly overexpressed in Breast by our test, that were also among class III genes related to breast cancer in Table 1 of [50], were FGFR1, BAG4, MDM2, YWHAB, ZNF217. For Lung, they were EGFR, MET, YWHAZ, MYC, NKX2-1, DCUN1D1. These findings would require further confirmation

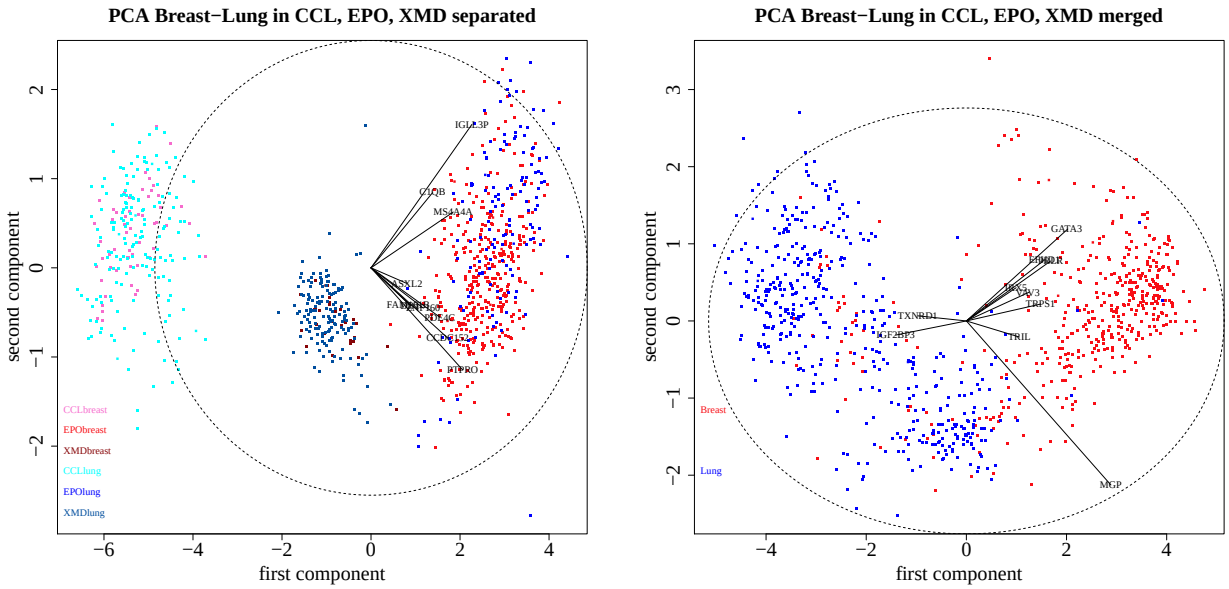

Figure 4: Principal component analysis of two assays of 10 symbols in 6 submatrices, extracted from CCL, EPO, and XMD according to keywords "Breast" and "Lung". The six submatrices are separated on the right panel, they have been merged on the right panel. In each case the 10 most significant features for the Kruskal-Wallis test are taken as variables. The two sets of 10 symbols are disjoint. Samples are represented by red points (Breast) or blue points (Lung).

over larger datasets. Yet they provide evidence that meaningful biological information can be extracted by merging generalist matrices such as CCL, EPO and XMD.

Our next experiment consisted in merging the two groups of blood RNA datasets, found to be homogeneous on the correlation analysis (Figure 1): HAV, HPS, PLE, WBS for healthy individuals, AML, MDS, for leukemias. The samples of MIL were separated into MILh (Healthy), and MILl (leukemias). The left panel of Figure 3 shows the first plane of the PCA for the same 22 features as in Figure 2, the 8 matrices being represented by different colors. It turns out that the samples corresponding to MILh are mixed on the representation with the other MIL points. Thus they were removed from the matrix "Healthy", whereas "Leukemia" was made by merging AML, MDS, and MILl. The Kruskal-Wallis test between Healthy and Leukemia, detected 16 977 significant features out of 17 691, among which 7 970 had a null p-value. The right panel of Figure 5 shows the PCA over 10 of them.

The one-sided Wilcoxon test was run to detect which symbols were significantly overexpressed in leukemias. For that test, a set of 4 191 symbols had a null p-value. A second set of symbol was extracted from C2: those appearing in leukemia-related genesets. The C2 set has 5 688 symbols, and the intersection with the first contains 1 617, which is highly significant for Fisher's hypergeometric test ($P = 1.36 \times 10^{-51}$). The ten symbols found most significant for leukemia by the Wilcoxon test were RPL34, GABARAP, RPL36A, H2AFV, CSDE1, DNTTIP2, OPHN1, PABPC3, PNRC1, RPSA. Among those 10, 8 appeared in the leukemia-related genesets of C2. The symbol H2AFV is found in six of them. Another noteworthy result concerns the pair of genes NUP98-TOP1, shown to be related with leukemia in [51]. When genes are

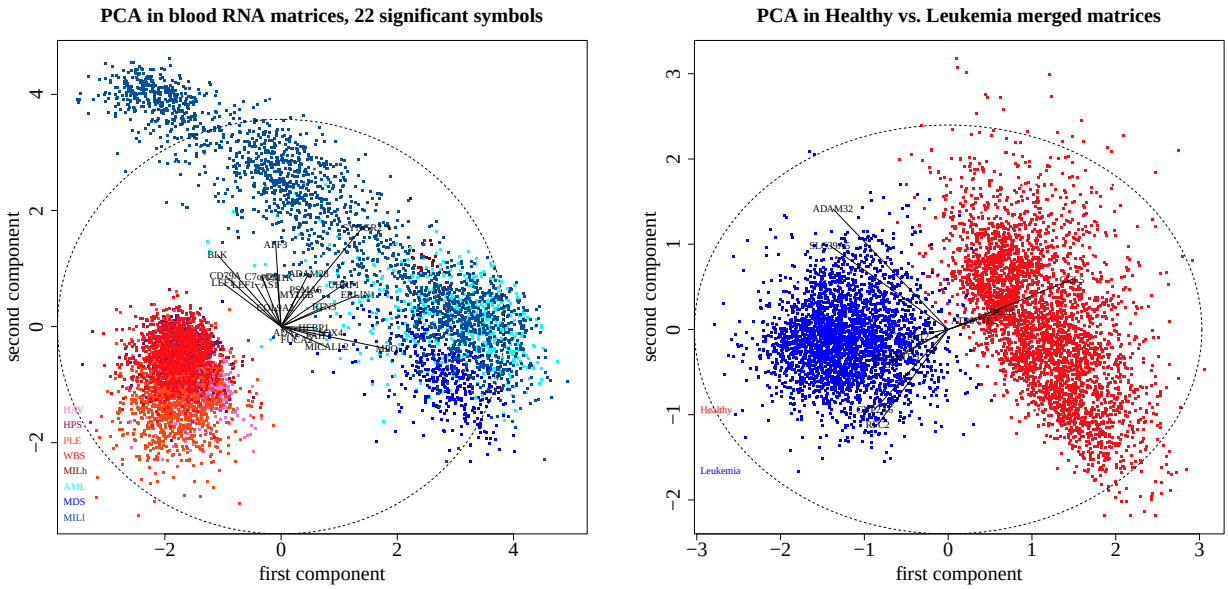

Figure 5: On the left panel, PCA of HAV, HPS, PLE, WBS, MILh (Healthy = red points), and AML, MDS, MILl (Leukemia = blue points), for the same 22 features as in Figure 2. On the right panel, HAV, HPS, ML, WBS, have been grouped into Healthy, AML, MDS, MILl into Leukemia. The assay is made of 10 random features among the 7 970 having a null Kruskal-Wallis p-value.

ranked by decreasing order of significance, NUP98 and TOP1 have ranks 187 and 65 respectively, which confirms their link with leukemia.

Another experiment was run on the same matrices, by separating acute myeloid leukemia samples, from all other samples. Thus the same calculations that had been run inside MIL before, were repeated over a larger number of samples. The acute myeloid leukemia samples were taken from AML and MIL (1 076 samples), others were obtained by merging HAV, HPS, PLE, WBS, MDS, with the non-AML samples of MIL (6 010 samples). The one-sided Wilcoxon test of comparison was run. For the 7 genes signaled as overexpressed in AML by [43], the results were much more significant as before: the least significant p-value was that of BIK: 8.4×10^{-35}, whereas FUT4 and HOMER3 had p-values below machine precision. Contrarily to the study that had been conducted inside MIL, a clear confirmation was also obtained for the genes reported by [43] to be underexpressed in case of AML. Five of them were in the common features of our matrices, four had p-values smaller than 10^{-100} for underexpression in AML. In particular, PELO and PLXNC1 who had not been found significantly underexpressed in the first experiment, now had p-values 3.5×10^{-238} and 4.4×10^{-168} in the test on merged matrices.

Discussion

A new set of R functions has been developed. Like other packages [15, 16], it performs the usual formatting operations. It also offers new functionalities for sorting lists of datasets according to information keywords. Various robust statistics techniques are encoded. The script and a user manual have been made available online. Using these R functions, a large scale study of 20 GEO datasets, totalling 17 745 samples, has been

conducted.

Our first conclusion is that Haibe-Kains et al. [4] were right in observing that inconsistencies between datasets make it dangerous to merge them without precautions. The risk is to declare as biologically significant, observations which are actually statistical artifacts. The first precaution is to transform the data into distribution-free values, i.e. to use robust statistics. This implies replacing the data of each sample by their empirical distribution function, or some other distribution-free score [7, 8]. Even after data have been homogenized, important discrepancies remain. For this reason, checking comparability between studies before merging them is imperative. One possible measure of similarity (among others, see for instance [42]) for two datasets is the correlation between medians, which has been used here. Two sets of samples corresponding to different conditions inside one given homogeneous dataset usually have correlations of medians above 0.8 (see Table 3). Arguably, it can be considered that two different datasets can safely be merged, if all paired-correlations between medians are above 0.8. This is not always the case, even between datasets coming from the same tissues, obtained under the same platform (see Table 2). Further ways of investigating possible discrepancies involve multivariate statistics. Graphical methods include Factor Analysis, Principal Component Analysis, Discriminant Analysis [52]. Inference can be done using the robust equivalents of usual normal-sample methods, i.e. Wilcoxon test instead of Student's t-test, Kruskal-Wallis instead of one-way anova, etc. When repeatedly applying such a test to a set of symbols, a False Detection Rate (FDR) correction must be applied to the p-values. We have chosen the Benjamini-Yekutieli method [21]. Our observation was that, even after FDR correction, the tests usually detect a sizeable proportion of all symbols as significant for discrimination, either between several different datasets, or between different types of samples within the same dataset. We believe that relevant biological information can be obtained from applying a discriminating test, then ranking features according to their degree of significance, i.e. ordering the values obtained over each feature by the test statistic. In the cases considered here (breast tumors against lung tumors, healthy blood samples against leukemias, acute myeloid leukemia against other blood RNA samples), it was observed that among the most significant symbols, a large proportion of them were already known as being related to the corresponding cancers. This can be viewed as evidence that meaningful biological information can be extracted by merging different datasets. We believe that important new findings could be obtained by the same method, being aware that a statistical listing of significant symbols does not necessarily imply that all listed symbols correspond to true biological information. Such a list must necessarily be expert-curated for biochemical validation.

Acknowledgements

BY, KC, and JJF acknowledge financial support from Laboratoire d'Excellence TOUCAN (Toulouse Cancer).

Additional Files

Additional material has been provided as a compressed directory available online:
`http://ljk.imag.fr/membres/Bernard.Ycart/publis/sagd.tgz`
It contains:

1. a R script file `sagd.r`: the R functions implementing the method described here,

2. a pdf file `sagd_manual.pdf`: user manual for the R functions.

References

1. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**:207–210.
2. Heider A, Alt R: **virtualArray: a R/bioconductor package to merge raw data from different microarray platforms**. *BMC Bioinformatics* 2013, **14**:75. [R package version 1.8.0].
3. Taminau J: *inSilicoMerging: Collection of Merging Techniques for Gene Expression Data* 2014, [http://insilicodb.com/]. [R package version 1.8.0].
4. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, Quackenbush J: **Inconsistency in large pharmacogenomic studies**. *Nature* 2013, **504**.
5. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies**. *Bioinformatics* 2002, **18**:405–412.
6. Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B, Costello CM, Schreiber S: **A comparison of oligonucleotide and cDNA-based microarray systems**. *Physiol Genomics* 2004, **16**:361–370.
7. Gibbons JD, Chakraborti S: *Nonparametric statistical inference*. Dekker, Basel, 4$^{\text{th}}$ edition 2003.
8. Héritier S, Cantoni E, Copt S, Victoria-Cantoni MP: *Robust methods in biostatistics*. Wiley, New York 2009.
9. Tsodikov A, Szabo A, Jones D: **Adjustments and measures of differential expression for microarray data**. *Bioinformatics* 2002, **18**:251–260.
10. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes**. *BMC Bioinformatics* 2005, **6**:265.
11. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments**. *FEBS Lett.* 2004, **573**:83–92.
12. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2013, [http://www.R-project.org/]. [ISBN 3-900051-07-0].
13. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, THothorn, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol.* 2004, **5**.
14. Gentleman R, Carey V, Huber W, Hahne F: *genefilter: genefilter: methods for filtering genes from microarray experiments*. [R package version 1.46.1].
15. Davis S, Meltzer P: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinformatics* 2007, **14**:1846–1847.
16. Taminau J: *inSilicoDb: Access to the InSilico Database* 2011, [https://insilicodb.org]. [R package version 1.7.4].
17. Carlson M: *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data(chip hgu133plus2)*. [R package version 2.8.0].
18. Carlson M: *hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)*. [R package version 2.8.0].
19. Dunning M, Lynch A, Eldridge M: *illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3)*. [R package version 1.16.0].
20. Dunning M, Lynch A, Eldridge M: *illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4)*. [R package version 1.16.0].
21. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency**. *Ann. Statist.* 2001, **29**:1165–1188.

22. **Expression Project for Oncology (expO)**[http://www.intgen.org/research-services/biobanking-experience/expo/].
23. Chen L, Wang S, Zhou Y, Wu X, et al.: **Identification of early growth response protein 1 (EGR-1) as a novel target for JUN-induced apoptosis in multiple myeloma**. *Blood* 2010, **115(1):**:61–70.
24. de Jonge HJ, Valk PJ, Veeger NJ, ter Elst A, et al.: **High VEGFC expression is associated with unique gene expression profiles and predicts adverse prognosis in pediatric and adult acute myeloid leukemia**. *Blood* 2010, **116(10)**:1747–54.
25. Roth R: **Human body index - transcriptional profiling**. *Gene Expression Omnibus (GEO)* NCBI2007:Series GSE7307.
26. Haferlach T, Kohlmann A, Wieczorek L, Basso G, et al.: **Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group**. *J Clin Oncol* 2010, **28(15)**:2529–37.
27. Mills KI, Kohlmann A, Williams PM, Wieczorek L, et al.: **Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome**. *Blood* 2009, **114(5)**:1063–72.
28. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, et al.: **Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA**. *PLoS Genet* 2011, **7(8)**:e1002197.
29. Popovici V, Chen W, Gallas BG, Hatzis C, et al.: **Effect of training-sample size and classification difficulty on the accuracy of genomic predictors**. *Breast Cancer Res* 2010, **12(1)**:R5.
30. Frei E, Visco C, Xu-Monette ZY, Dirnhofer S, et al.: **Addition of rituximab to chemotherapy overcomes the negative prognostic impact of cyclin E expression in diffuse large B-cell lymphoma**. *J Clin Pathol* 2013, **66(11)**:956–61.
31. Westra HJ, Peters MJ, Esko T, Yaghootkar H, et al.: **Systematic identification of trans eQTLs as putative drivers of known disease associations**. *Nat Genet* 2013, **45(10)**:1238–43.
32. Barretina J, Caponigro G, Stransky N, Venkatesan K, et al.: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature* 2012, **483(7391)**:603–7.
33. Hernandez DG, Nalls MA, Moore M, Chong S, et al.: **Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain**. *Neurobiol Dis* 2012, **47(1)**:20–8.
34. Mayerle J, den Hoed CM, Schurmann C, Stolk L, et al.: **Identification of genetic loci associated with Helicobacter pylori serologic status**. *JAMA* 2013, **309(18)**:1912–20.
35. Xiao W, Mindrinos MN, Seok J, Cuschieri J, et al.: **A genomic storm in critically injured humans**. *J Exp Med* 2011, **208(13)**:2581–90.
36. Seok J, Warren HS, Cuenca AG, Mindrinos MN, et al.: **Genomic responses in mouse models poorly mimic human inflammatory diseases**. *Proc Natl Acad Sci U S A* 2013, **110(9)**:3507–12.
37. Marisa L, de Reyniès A, Duval A, Selves J, et al.: **Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value**. *PLoS Med* 2013, **10(5)**:e1001453.
38. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, et al.: **Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association**. *Hum Mol Genet* 2011, **20(20)**:4082–92.
39. Esko T, Metspalu A: **Gene Expression profiling in healthy population samples**. *Gene Expression Omnibus (GEO)* NCBI2013:Series GSE48348.
40. Hollingshead MG, Stockwin LH, Alcoser SY, Newton DL, et al.: **Microarray analysis of xenograft models in use at the Developmental Therapeutics Program of the National Cancer Institute (DTP-NCI)**. *Gene Expression Omnibus (GEO)* NCBI2013:Series GSE48433.
41. Obermoser G, Presnell S, Domico K, Xu H, et al.: **Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines**. *Immunity* 2013, **38(4)**:831–44.

42. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E: **A Cross-Study Comparison of Gene Expression Studies for the Molecular Classification of Lung Cancer**. *Clin Cancer Res* 2004.

43. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, et al.: **identification of genes with abnormal expression changes in acute myeloid leukemia**. *Genes Chromosomes Cancer* 2008, **47**:8–20.

44. Zakaria Z, et al.: **Identification of Estrogen-Related Genes in Breast Cancer: The Malaysian Context**. *The Open Breast Cancer Journal* 2010, **2**:16–24.

45. Ma CX, Ellis MJ: **The Cancer Genome Atlas: Clinical Applications for Breast Cancer**. *Oncology* 2013, **27(12)**:1263–9.

46. Beljan PR, Durdov MG, Capkun V, Ivcevic V, Pavlovic A, Soljic V, Peric M: **IMP3 can predict aggressive behaviour of lung adenocarcinoma**. *Diagn Pathol.* 2012, **7**:165.

47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS* 2005, **102**:15545–15550.

48. Mosca E, Alfieri R, Merelli I, Viti F, Calabria A, Milanesi L: **A multilevel data integration resource for breast cancer study**. *BMC Syst Biol.* 2010, **4**:76.

49. Marchitti S, Orlicky D, Brocker C, Vasiliou V: **Aldehyde Dehydrogenase 3B1 (ALDH3B1): Immunohistochemical Tissue Distribution and Cellular-specific Localization in Normal and Cancerous Human Tissues**. *J Histochem Cytochem* 2010, **58(9)**:765–783.

50. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS: **A census of overexpressed human cancer genes**. *Nature Reviews Cancer* 2010, **10**:59–64.

51. Gurevich RM, Aplan PD, Humphries RK: **NUP98-Topoisomerase I acute myeloid leukemia-associated fusion gene has potent leukemogenic activities independent of an engineered catalytic site mutation**. *Blood* 2004, **104**:1127–1136.

52. Härdle WK, Simar L: *Applied Multivariate Statistical Analysis*. Springer, New York 2012.