Electronic Supplementary Materials (ESI)

Q-GDEMAR: A METHOD FOR THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES IN MICROARRAYS WITH UNBALANCED GROUPS

Daniel V. Guebel[1,2], Montserrat Perera-Alberto[1,3], Néstor V. Torres[1*]

[1]Systems Biology and Mathematical Modelling Group. Science Faculty. University of La Laguna. Tenerife. Canary Islands. Spain.

[2] Biotechnology Counselling Services. Buenos Aires, Argentina.

[3]. Department of Anatomy, Pathology, Histology and Physiology. University of La Laguna. Tenerife. Canary Islands. Spain.

* e-mail: (NVT) ntorres@ull.edu.es

Annex 1. Computation of p-values

a. Apply a $\log_2$-transformation to the matrix of microarray standardized data.

b. Compute the values of the discriminating variable chosen (see Section 2.3).

c. Determine the *quantiles* (**Q**) associated to the following series of cumulated relative frequencies: $1x10^{-6}$, $1x10^{-5}$, $1x10^{-4}$, $1x10^{-3}$, $1x10^{-2}$, 0.25, 0.5, 0.75, 0.99, 0.999, 0.9999, 0.99999, 0.999999). To this aim we used the program Matlab v.7.10 (Mathworks, USA). Although “relative frequencies” and “probabilities” are different mathematical entities (i.e., probabilities are the asymptotic limit of the relative frequencies when the number of replications tend to infinity), herein we treat both concepts as equivalents given the great number of genes tested in the microarray. Note that quantiles $Q_{0.25}$ and $Q_{0.75}$ are of interest to determine the Inter-Quartile Range (IQR; $IQR=Q_{0.75}-Q_{0.25}$). The quantile $Q_{0.50}$ is the median value of the distribution.

d. Once the quantiles corresponding to the tails ($Q_{1x10}{}^{-6}$ to $Q_{1x10}{}^{-2}$ and $Q_{0.99}$ to $Q_{0.999999}$) are known, determine the best mathematical function that fits the probabilities **p** as a function of the quantile values (**Q**). Note that each tail has to be fitted separately. We used the CurveExpert program (SWREG digital river, USA) for this purpose. This software performs the process automatically by testing an ample menu of built-in fitting functions. Thus, a very high R^2 coefficient for the fitting is obtained. However, if a dedicated software is not available, the fitting of the tails can still be achieved by testing some simple functions, such as exponential ($y=a.e^{b.x}$), potential ($y=a.x^b$), and quadratic inverse ($y=1/(a+b.x+c.x^2)$).

e. Return to the original data sheet containing the values obtained in point *b*. Sort the data sequentially both in descending and ascending order according to the value of the *discriminating variable*. Select the rows comprised between the maximum (or minimum value as corresponds) of the discriminating variable and the row in which the discriminating variable reaches a value higher than the quantile value associated to p=0.99 (or p=0.01 as corresponds). Copy these data to a new sheet together with the names of the associated probes and genes.

f. Compute the associated probability in a descending-column order by introducing the fitting function previously determined in point (d). Now you have a presumptive list of up- and down-regulated genes and their associated p-values.

g. To get the final list of differentially expressed genes, filter the list obtained in the previous point according to the FDR criterion (see Annex 2).

Annex 2. Computation of the false discovery rate (FDR)

FDR is defined as the “expected” value of the quotient between the number of false positive cases and the number of cases declared as significant.[1-4] In our approach, the number of false positive cases is calculated by taking into consideration that the central region of the data always follows a Gaussian distribution (Figure 2.A, Section 2.2.1). From this central region it is then possible to estimate its extreme values. These extreme values can then be used as cut-offs of the tails at the entire data distribution, thus allowing us to compute the number of genes declared significant in each tail (Figure 2.B, Section 2.2.1), and hence, finally compute the FDR (Figure 3, Section 2.2.2).

In brief:

a. Compute the standard deviation (SD) for the Gaussian central region of the data distribution by using equation 7.

$$SD_{\text{Gaussian central region}} = (1/2)(Q_{0.75}-Q_{0.25})/0.6745 \qquad (7)$$

Based on the property of symmetry, the $SD_{\text{Gaussian central region}}$ was obtained in equation 7 by averaging two available estimations of it, one from the right-side (SD_1) and the other from the left-side (SD_2). Both estimations are derived from the well-established relations present in Gaussian distributions, such that $Q_{0.75}=Q_{05}+0.6745SD_1$ and $Q_{0.25}=Q_{0.5}-0.6745SD_2$ respectively. In addition, note that the mean value of the distribution has been substituted by the quantile Q_{050} (i.e., the median value). The factor 0.6745 was taken from a normal z-probabilities table.

b. Determine the expected extreme values of the discriminating variable for the genes which do not change their expression significantly (i.e., those belonging to the Gaussian central region) by using the equations 8 and 9. Taking a value of z= 3.09 is usually enough as it yields a value of p=0.001 at each tail of the distribution.

$$L_u = Upper\ Limit = Q_{0.5} + z\ SD_{Gaussian\ central\_region}$$
(8)

$$L_L = Lower\ Limit = \ Q_{0.5} - \ z\ SD_{Gaussian\ central\_region}$$
(9)

c. Count the number of genes declared as significant. To this aim, count the cases in which the values of the discriminating variable exceed the allowed upper limit of the central region (n^+= cases where discriminating variable value > Lu). Also count the number of genes in which the discriminating variable is lower than the lower limit of the central region (n^- = cases where discriminating variable value < L_L). This operation can be done by simple inspection of the data sheet.

d. Compute the total number of genes declared significant (n*), by performing the following summation:

$$n^* = n^+ + n^-$$
(10)

e. Compute the expected number of false positive genes:

$$N^o Expected\ False(+)_{(right\ tail)} = N^o Expected\ False(+)_{(left\ tail)} =$$

$$\frac{1}{2}\ N^o Genes_{(Gaussian\ central\ region)} 0.001$$
(11)

In turn, the number of genes comprised in the Gaussian central region that is required in equation 11 is calculated as follows:

$$N^o genes_{(Gaussian\ central\ region)} = N^o genes_{(microarray)} - (n^+ + n^-)$$
(12)

Importantly, according to the Gaussian model the number of false positives will always be lower or equal to the number declared significant from the tails of the entire empirical distribution (0≤FDR ≤1).

f. Finally, the percent of false discovery rate (%FDR) can be computed as follows:

$$\%FDR_{(upper-regulated\ genes)} = \frac{N^o Expected\ False\ (+)_{(right\ tail)}}{N^+}\ 100$$
(13)

$$\%FDR_{(down-regulated\ genes)} = \frac{N^o Expected\ False\ (+)_{(left\ tail)}}{N^-}\ 100$$
(14)

g. In exceptional cases, when the FDR obtained is not satisfactory, it is necessary to return to equations 8 and 9. Modify the z-value in them so that its associated probability becomes lower than 1x10^{-3}. For example, if z=3.891 this will result in a p-value=5x10^{-5} at each tail of the Gaussian central distribution. Now repeat the computation procedure to obtain the new FDR values.
h. If the FDR values in equations 13 and 14 are satisfactory, the procedure is stopped, and two FDR values will be available, one for each tail of the data distribution.

Note: As general procedure, we have established that it is necessary to choose an starting value for probability p=(1-alpha error) and apply iteratively our algorithm until to arrive to an acceptable value of FDR (Figure 3, Section 2.2.2). In practice, we recommended the use of two particular z-values (z=3.09 and z=3.89), which must be tested in sequential order. As was stated previously, the value of z=3.09 is associated to a p-value of 1x10^{-3}, while z=3.89 provides a p-value of 5x10^{-5}. By using these two herein suggested z-values, we have verified that FDR in our approach showed an excellent performance (median value=0.67%, interquartile range IQR = 0.87%; Section 3.2). Instead, in a quarter of the microarrays tested, the FDR computed by LIMMA, applying Empirical Bayes and Benjamini-Hochberg, FDR ranged between 17 to 45% (Section 3.2) while in another quarter of the microarrays tested by LIMMA, the FDR was as high (~85-90%) that not allowed the detection of any gene, while we could detect (Section 3.2, and Table S1 in ESI). In fact, The FDR performance in our approach using the suggested z-values was associated to an excellent sensitivity in the detection of differential expressed genes (Section 3.2 and Section 3.4).

However, there is no problem if it is wished to assay other possible z-values. For example, in-between the suggested z-values it can tested a value of z=3.291 which is associated to p-value= 5x10^{-4}. Instead, if it is wished or needed to test a z-value higher than z=3.89, it can attempt with a value of z=4.417, which is associated to p-value=5x10^{-6}. All the p-values herein given were taken from a Z-normal table, and correspond to one-side tail. In any case, each one of the z-value considered, must be applied to equations 8-9, and then follow the algorithm through equations 10-14 until to arrive to the corresponding FDR values. This cycle must be repeated iteratively for each one of the z-values tested until to arrive to an acceptable FDR value (Figure 3, Section 2.2.2).

References

1. Y. Benjamini, Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Statistical Society*, Series B (Methodological), 1995, **57**: 289-300.
2. B. Efron, R. Tibshirani, J. D. Storey, V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *J American Statistical Association*, 2001, **96**, 1151-1160.
3. K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald et al. Comparison of seven methods for producing Affymetrix expression scores

based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*, 2005, 6**:** 26.

4. E. Hansen, K. F. Kerr. A comparison of two classes of methods for estimating false discovery rates in microarray studies. *Scientifica*, 2012, article ID: 519394. Available at: http://www.hindawi.com/journals/scientifica/2012/519394/abs/

Table S1. Comparison of the Q-GDEMAR performance against the results obtained by the reference LIMMA. The number of deregulated genes detected is compared at a similar level of FDR. Calculations were done after the $\log_2$-transformation. The FDR values are indicated between parentheses (as percent).

Microarray data set	Method	Up-regulated Genes	Down-regulated genes
GSE35713 (44 control samples vs. 11 treatment samples)[a]	Δ1 Difference	242 *(0.69%)*	638 *(0.19%)*
	LIMMA	**0** *(0.69%)*	**1** *(0.19%)*
	Median Ratio	169 *(1.23%)*	398 *(0.52%)*
	LIMMA	163 *(1.23%)*	**6** *(0.52%)*
	Δ2 Difference	456 *(4.58%)*	840 *(2.44%)*
	LIMMA	650 *(4.58%)*	**33** *(2.44%)*
GSE36297 (10 control samples vs. 6 treatment samples)[b]	Median Ratio	776 *(0.44%)*	558 *(2.98%)*
	LIMMA	**0** *(0.44%)*	**0** *(2.98%)*
	Δ2 Difference	108(0.76%)	120(1.14%)
	LIMMA	**0** *(0.76%)*	**0** *(1.14%)*
	Δ1 Difference	172 *(0.60%)*	104 *(0.99%)*
	LIMMA	**0** *(0.60%)*	**0** *(0.99%)*
GSE48754 (5 control samples vs. 3 treatment samples)[c]	Δ2 Difference	538 *(0.39%)*	877 *(0.02%)*
	LIMMA	**17** *(0.39%)*	**0**(0.02%)
	Median Ratio	477 *(0.46%)*	345 *(0.64%)*
	LIMMA	**17** *(0.46%)*	**1** *(0.63%)*
	Δ1 Difference	292 *(0.74%)*	263 *(0.82%)*
	LIMMA	**19** *(0.74%)*	**1** *(0.82%)*
GSE5281 (10 control samples vs. 3 treatment samples)[d]	Median Ratio	2845 *(0.66%)*	3067 *(0.61%)*
	Δ1 Difference	**911** *(2.35%)*	**1** *(21.4%)*
	Δ2 Difference	686 *(3.05%)*	1211 *(1.7%)*
	LIMMA[(*)]	**0** *(99.6%)*	**0** *(99.6%)*
GSE54992 (6 control samples vs. 9 treatment samples)[e]	Δ1 Difference	1368 *(1.48%)*	1745 *(1.16%)*
	LIMMA	3145 *(1.48%)*	2506 *(1.16%)*
	Δ2 Difference	737 *(2.85%)*	826 *(2.54%)*
	LIMMA	3830 *(2.85%)*	3226 *(2.54%)*
	Median Ratio	2436 *(0.81%)*	1663 *(1.19%)*
	LIMMA	2591 *(0.81%)*	2529 *(1.19%)*

GSE48060 (21control samples vs. 31 treatment samples)[f]	Δ2 Difference	482 *(0.42%)*	1605 *(0.12%)*
	LIMMA	**1** *(0.42%)*	**0** *(0.12%)*
	Median Ratio	141 *(1.48%)*	405 *(0.52%)*
	LIMMA	**8** *(1.48%)*	**1** *(0.52%)*
	Δ1 Difference	24 *(7.54%)*	735 *(2.39%)*
	LIMMA	86 *(7.54%)*	**0** *(2.39%)*
GSE28619 (7 control samples vs. 15 treatment samples)[g]	Δ2 Difference	7496*(0.39%)*	5481*(0.54%)*
	LIMMA	**3378** *(0.39%)*	**1128***(0.54%)*
	Median Ratio	6616*(0.47%)*	5034*(0.68%)*
	LIMMA	3526*(0.47%)*	2306*(0.68%)*
	Δ1 Difference	3705*(0.96%)*	3312(*1,07%)*
	LIMMA	4058*(0.96%)*	2631*(1,07%)*
GSE1919 (5 control samples vs. 5 treatment samples)[h]	Median Ratio	949 *(0.55%)*	645 *(0.82%)*
	LIMMA	**56** *(0.55%)*	**50** *(0.82%)*
	Δ1 Difference	800 *(0.67%)*	554 *(0.97%)*
	LIMMA	**59** *(0.67%)*	**56** *(0.97%)*
	Δ2 Difference	238 *(2.44%)*	295 *(1.97%)*
	LIMMA	**167** *(2.44%)*	**100** *(1.97%)*
GSE34308 (5 control samples vs. 5 treatment samples)[i]	Δ2 Difference	1254 *(0.16%)*	592*(0.34%)*
	LIMMA	**0***(0.16%)*	**0**(0.34%)
	Median Ratio	849 *(0.24%)*	226*(0.91%)*
	LIMMA	**0***(0.24%)*	**0***(0.91%)*
	Δ1 Difference	523*(0.39%)*	212*(0.98%)*
	LIMMA	**0***(0.39%)*	**0***(0.98%)*
GSE1297 (7 control samples vs. 2 treatment samples)[j]	Median Ratio	453*(0.22%)*	541*(0.18%)*
	LIMMA	**0***(0.22%)*	**0***(0.18%)*
	Δ2 Difference	541*(0.18%)*	135*(0.77%)*
	LIMMA	**0***(0.18%)*	**0***(0.77%)*
GSE11882 (10 control samples vs 11 treatment samples)[k]	Median Ratio	2244*(0.09%)*	641*(0.32%)*
	LIMMA	**0***(0.09%)*	**0***(0.32%)*
	Δ2 Difference	2640*(0.75%)*	1315*(0.15%)*
	LIMMA	**0***(0.75%)*	**0***(0.15%)*
	Δ1 Difference	2225*(0.09%)*	596*(0.34%)*
	LIMMA	**0***(0.09%)*	**0***(0.34%)*
GSE46922 (10 control samples vs 11 treatment samples)[l]	Median Ratio	3519(0.05%)	620(0.34%)
	LIMMA	**0***(0.05%)*	**0***(0.34%)*
	Δ2 Difference	852*(0.25%)*	1315*(0.15%)*
	LIMMA	**0***(0.25%)*	**0***(0.15%)*
	Δ1 Difference	1330*(0.15%)*	2667*(0.07%)*
	LIMMA	**0***(0.15%)*	**0***(0.07%)*

Across of the analysed microarrays, control and treatment samples correspond to the following conditions: [a]Healthy vs. patients with long-time diabetes type 1; [b]Normal vs. patients with mutation in insulin receptor; [c]Normal vs. patients with Swedish myopathy; [d]Male vs. Female in healthy ancient; [e]Healthy vs. patients with tuberculosis; [f]Normal vs. patients with first-time myocardial infarct; [g]Healthy vs. patients with alcoholic hepatitis; [h]Healthy vs. patients with rheumatoid arthritis; [i]Healthy vs. patients with childhood cerebral form of X-linked adreno-leuko dystrophy; [j,k]Male vs. Female in healthy acient; [l]acute vs. chronic immune thrombocytopenia.

[*] In this case, LIMMA did not allow to detect any significant gene due to the high value of FDR (99.6%) computed for all the genes, whichever were their $\log_2$(FC) values.

Table 2S: Efficiency of the different Q-GDEMAR variants and shape characterization of the distributions. The different measurements of the variant's efficiency are computed on the basis of results in Table S1. The parameters of the distributions are computed over each microarray data-set, using a given discriminant variant on the $\log_2$-transformed values down-loaded from GEO database.

Microarray Data // Variant Methods	**Data Distribution Parameters**		**Variant's Efficiency**		
	Skewness Coefficient	**Kurtosis Coefficient**	**Φ index** [a]	**Total Significant Genes Detected**	$\overline{FDR}$ **(%)**
		GSE35713			
Δ1-difference	-0.5	6	0.12	880	0.33
Median Ratio	-0.3	7	0.23	567	0.73
Δ2-difference	-0.7	11	0.22	1296	3.16
		GSE36297			
Median Ratio	0.7	28.7	0.57	2442	0.07
Δ1-difference	0.4	10.5	0.21	276	0.75
Δ2-difference	-0.04	19.2	0.20	228	0.96
		GSE48754			
Δ2-difference	-0.2	21	0.40	1415	0.16
Median Ratio	1.6	31	0.31	822	0.53
Δ1-difference	0.4	14.5	0.21	555	0.78
		GSE5281			
Δ2-difference	2.4	27	0.55	5912	0.63
Median Ratio	-0.2	5	0.26	1897	2.12
Δ1-difference	1.1	5	0.11	912	2.37
		GSE54992			
Median Ratio	173	33790	0.89	4225	0.93
Δ1-difference	-0.5	9.4	0.39	1745	1.30
Δ2-difference	-0.3	7.7	0.25	1563	2.68
		GSE48060			
Δ2-difference	1.0	65	0.21	2087	0.24
Δ1-difference	0.7	6	0.12	759	2.55

		GSE28619			
Median Ratio	4.8	72.5	0.95	11050	0.56
Δ2-difference	0.3	23.3	0.90	12977	0.33
Δ1-difference	-0.3	17.3	0.67	7017	1.01
		GSE1919			
Median Ratio	4	236	0.89	949	0.66
Δ1-difference	31	3100	0.97	1354	0.79
Δ2-difference	0.4	11	0.92	533	2.18
		GSE34308			
Δ2-difference	0.9	27.6	0.46	1846	0.22
Median Ratio	3.1	46.3	0.37	1075	0.38
Δ1-difference	0.8	13.6	0.27	724	0.58
		GSE1297			
Median Ratio	1.9	41.1	0.45	1572	2.44
Δ2-difference	-0.21	7.3	0.23	237	0.87
		GSE11882			
Median Ratio	147.2	24100	0.81	2885	0.14
Δ2-difference	79.5	7200	0.92	3955	0.55
Δ1-difference	1.2	18.8	0.48	2821	0.14
		GSE46922			
Median Ratio	6.2	870	0.91	6669	0.005
Δ1-difference	-0.8	8.8	0.56	3997	0.09
Δ2-difference	-0.2	5.9	0.11	1472	0.29

[a] $\Phi \text{ index} = \left(\mathbf{1} - \frac{\sigma_{central}}{\sigma_{global}}\right)$, where σ_{central} is the standard deviation of the Gaussian central region, and σ_{global} is the standard deviation of the entire data distribution.

Table 3S: Matrix of correlation between some intrinsic characteristics of the data distributions (Skewness and Kurtosis) and several output measurements (Φ index, Total genes detected,$\overline{\text{FDR}}$) computed upon the Q-GDEMAR variants shown in Table S2. Data were restricted to those belonging to a unique sub-set (HG-U133Plus2, Affymetrix). Pearson correlation coefficients (R) are displayed in the upper triangular matrix, while their significance probabilities are in the lower triangular matrix.

(n=23)	Skewness	Kurtosis	Φ index	Genes detected	$\overline{\text{FDR}}$ (%)
Skewness	1	0.6242**	0.5770**	0.2427	-0.1887
Kurtosis	*0.0015*	1	0.6249**	0.3224	-0.0880
Φ index	*0.0039*	*0.0014*	1	0.8060**	-0.3224
Genes detected	0.2645	0.1335	*0.0000*	1	-0.2492
$\overline{\text{FDR}}$ (%)	0.3884	0.6985	0.1335	*0.2515*	1

In Table 3S, note the occurrence of four very significant correlations (**, $p \leq 0.01$) despite the moderate values of the R coefficients. Importantly, the "total number of genes detected" correlates strongly with "Φ index" (R=0.8060), and this correlation persists even after being corrected by kurtosis (partial correlation $R_{genes\_\Phi(kurtosis)}$=0.8408) or skewness (partial correlation $R_{genes\_\Phi.(skewness)}$=0.8194). On the other hand, the remaining variables, when paired, only yield low partial correlation coefficients (data not shown). All together this means that some co-linear relations are present in the data. For this reason, Table S3 was subjected to principal component analysis (see Figure 7, main text).