INFLUENCE OF DATA NORMALIZATION METHOD IN THE ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES WITH MICROARRAYS

Barbara Gerič STARE[1] , Peter DOLNIČAR[2], Irena MAVRIČ PLEŠKO[3], Vladimir MEGLIČ[4]

[1] Kmetijski inštitut Slovenije, Oddelek za varstvo rastlin, Ljubljana
[2] Kmetijski inštitut Slovenije, Oddelek za poljedelstvo in semenarstvo, Ljubljana

ABSTRACT

DNA microarrays, a collection of microscopic DNA spots attached to a solid surface, can be used to simultaneously measure the expression levels of large numbers of genes. We have used the Potato Gene Expression Microarray produced by Agilent (POCI microarray) to determine differentially expressed (DE) genes in potato tubers due to infection with Potato virus Y (PVY). The POCI microarray is currently the most complete representation of the potato transcriptome and widely used in the potato research community. There is no universal consensus method on how to analyse the microarray data. In this work we have assessed influence of data normalization method in the analysis of differentially expressed genes with microarrays. Normalization enables correction of systematic differences between samples, which do not represent true biological variation between samples. Three selected methods for data normalisation between microarrays; quantile normalization, scale normalization and variance stabilizing normalization resulted in 190, 151 and 77 DE genes between 56 biological samples (36 infected with PVY and 20 uninfected), respectively. Among these, 67 genes were determined as DE with all three methods, while further 78 genes were determined as DE with quantile and scale normalization methods. There was in general a good coincidence of genes determined as DE genes with all three methods. However, different normalisation methods revealed considerably different number of DE genes.

Key words: data normalization, microarray, potato, potato virus Y, PVY, Solanum tuberosum

VPLIV IZBIRE METODE NORMALIZACIJE PODATKOV PRI ANALIZI RAZLIČNO IZRAŽENIH GENOV Z MIKROMREŽAMI

IZVLEČEK

DNA mikromreže so zbirka različnih molekul DNA vezanih na mikroskopsko majhna mesta na trdni podlagi, ki se lahko uporablja za hkratno merjenje nivoja izražanja velikega števila genov. Mikromreže za proučevanje izražanja genov krompirja, ki jih proizvaja podjetje Agilent (POCI mikromreže), smo uporabili za določitev nivoja izražanja genov v gomoljih krompirja okuženega s krompirjevim virusom Y (*Potato virus Y* - PVY). POCI mikromreža trenutno predstavlja najpopolnejši transkriptom krompirja in se v zadnjih letih v veliki meri uporablja v raziskavah krompirja. Za analizo podatkov pridobljenih z mikromrežami ne obstaja ena sama splošno sprejeta metoda. V tem delu smo ocenili vpliv metode normalizacije podatkov pri analizi različno izraženih genov z mikromrežami. Normalizacija omogoča popravek sistematičnih razlik med vzorci, ki ne predstavljajo resnične biotične razlike med vzorci. Tri izbrane metode za normalizacijo podatkov med mikromrežami; kvartilna normalizacija,

[1] dr., univ. dipl. biol., Hacquetova 17, SI-1000 Ljubljana
[2] mag., univ. dipl. inž. agr., prav tam
[3] dr., univ. dipl. biol., prav tam
[4] dr., univ. dipl. inž. agr., prav tam

normalizacija z uravnoteženjem in normalizacija s stabilizacijo variance so določile 190, 151 oz. 77 genov kot različno izraženih pri analizi 56 biotičnih vzorcev (36 okuženih s PVY in 20 neokuženih). Med njimi je bilo 67 genov določenih z vsemi tremi metodami, medtem ko je bilo dodatnih 78 genov določenih s kvartilno normalizacijo in normalizacijo z uravnoteženjem. Pokazali smo dobro skladnost pri določitvi različno izraženih genov z vsemi tremi metodami, čeprav smo z različnimi metodami določili občutno različno število genov.

Ključne besede: krompir, krompirjev virus Y, mikromreže, normalizacija podatkov, PVY, *Solanum tuberosum*

1 INTRODUCTION

DNA microarrays, a collection of microscopic DNA spots attached to a solid surface, can be used to measure the expression levels of large numbers of genes simultaneously. Process of microarray measurement consists of several steps, from determining the biological question, to the experimental design, carrying out the wet part of microarray experiment, image analysis, data normalization, data analysis of differentially expressed genes, biological verification and interpretation. Each step is presenting researchers with different changes, options and decisions that need to be made. Further, there is no universal consensus method on how to analyse the microarray data. Here we will concentrate only on data normalization methods and some aspects of experimental design.

Normalisation enables correction of systematic differences between samples, which do not represent true biological variation between samples. Normalisation between arrays normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays. There are several methods for data normalization between arrays. Quantile normalization was proposed by Bolstad *et al*. (2003) for Affymetrix-style single-channel arrays and by Yang and Thorne (2003) for two-color cDNA arrays. Quantile normalization ensures that the intensities have the same empirical distribution across arrays and across channels. The scale normalization method was proposed by Yang *et al*. (2001, 2002) and is further explained by Smyth and Speed (2003). The idea is simply to scale the log-ratios to have the same median-absolute-deviation (MAD) across arrays. The Variance Stabilizing Normalization (vsn) algorithm performs background correction and normalization simultaneously (Huber *et al*., 2002).

In this work we have assessed the influence of data normalization method in the analysis of differentially expressed genes with microarrays. We have used the Potato Gene Expression Microarray produced by Agilent (also known as POCI microarray) to determine differentially expressed (DE) genes in potato tubers due to infection with *Potato virus Y* (PVY). The POCI microarray is currently the most complete representation of the potato transcriptome and widely used in the potato research community.

2 MATERIAL AND METHODS

Potato tuber production and storage experiment of potato tubers at different temperature regimes have been described previously (Dolničar *et al*., 2011). Total RNA from tuber tissue was extracted with RNeasy Plant Mini kit (Qiagen). RNA quality and quantity were determined with 2100 Bioanalyser (Agilent Technologies) and NanoDrop ND-1000 UV-VIS spectrophotometer (ThermoScientific). The spiked total RNA was reverse transcribed into cDNA and then converted into labeled cRNA by in-vitro transcription with Quick Amp Labeling Kit One-Color (Agilent Technologies) in-corporating Cyanine-3-CTP. Cyanine-3-labeled cRNA samples were fragmented and prepared for One-Color based hybridization (Gene Expression Hybridization Kit, Agilent Technologies). Samples were hybridized at 65 ^{o}C for 17 hrs on

Potato Oligo Chip Initiative (POCI) Microarrays (4x44K format, AMADID 015425, Agilent). The microarrays were washed with increasing stringency using Gene Expression Wash Buffers (Agilent Technologies) followed by drying with acetonitrile (SIGMA). Fluorescent signal intensities were detected with Scan Control 8.4.1 Software (Agilent Technologies) on the Agilent DNA Microarray Scanner and extracted from the images using Feature Extraction 10.5.1.1 Software (Agilent Technologies) and the design file 015425_D_F_20061105.xml.
Within-array normalization and background correction were performed by Feature Extraction 10.5.1.1 Software (Agilent Technologies). Raw data of the samples was analysed in R (v. 2.12.2) statistical environment using Bioconductor package (v. 2.7) the limma package (Linear Models for Microarray Analysis). The data was analysed using single emission channel.
Three methods for data normalisation between microarrays (quantile normalization, scale normalization and variance stabilizing normalization) were applied to 56 biological samples representing 36 tubers infected with PVY and 20 uninfected tubers. Data was fit to the linear model and empirical Bayes statistics for differential expression was applied. Genes were determined as DE when B-value was above 4.0, adjusted p-value below 0.001, and lfc- value (log Fold Change = difference in expression level of a certain gene between two treatments) was at least 1 or above. Venn diagram was constructed with the on-line tool Make Venn Diagrams by Max-Planck Institut für Molekulare Pflantzenphysiologie (http://mapman.mpimp-golm.mpg.de/general/venn/index.html). To draw conclusions about the biological meaning of DE genes detected with microarrays, we visualized the data based on clustering of the genes to metabolic pathways and processes in the cell with the tool MapMan and mapping GoMapMan, which includes ontology for gene functions (www.gomapman.org).

3 RESULTS AND DISCUSSION

Normalisation between arrays normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays (Figure 1).
Three selected methods for data normalisation between microarrays implemented in limma package; quantile normalization, scale normalization and variance stabilizing normalization resulted in 190, 151 and 77 DE genes between 56 biological samples (36 infected with PVY and 20 uninfected), respectively. Among these, 67 genes were determined as DE with all three methods, while further 78 genes were determined as DE with quantile and scale normalization methods. Venn diagram illustrates relations of sets of DE genes determined with three different normalisation methods (Figure 2).
Different normalisation methods revealed different number of DE genes. There was a good coincidence of genes determined as DE genes with quantile and scale normalization. The variance stabilizing normalization resulted in the smallest number of DE genes. The reason for this underestimation of DE genes may lay in the fact that the vsn algorithm performs not only normalization but simultaneously also a background correction. However, the background correction was already performed on our data automatically at the step of image acquisition. Therefore we conclude that the variance stabilizing normalization is not a good choice of normalisation method for our data.
Every spot on the POCI array representing particular gene has assigned a description of the biological function (BIN). The three different normalisation methods: quantile, scale and variance stabilizing normalization have determined DE genes from 18, 15 and 13 BINs respectively. Identical BINs were determined with all three methods and some additional BINs were determined with the first two methods (Figure 3, Table 1). BIN 20 comprising stress related genes was determined with quantile normalization and scale normalization method but not with variance stabilizing normalization.

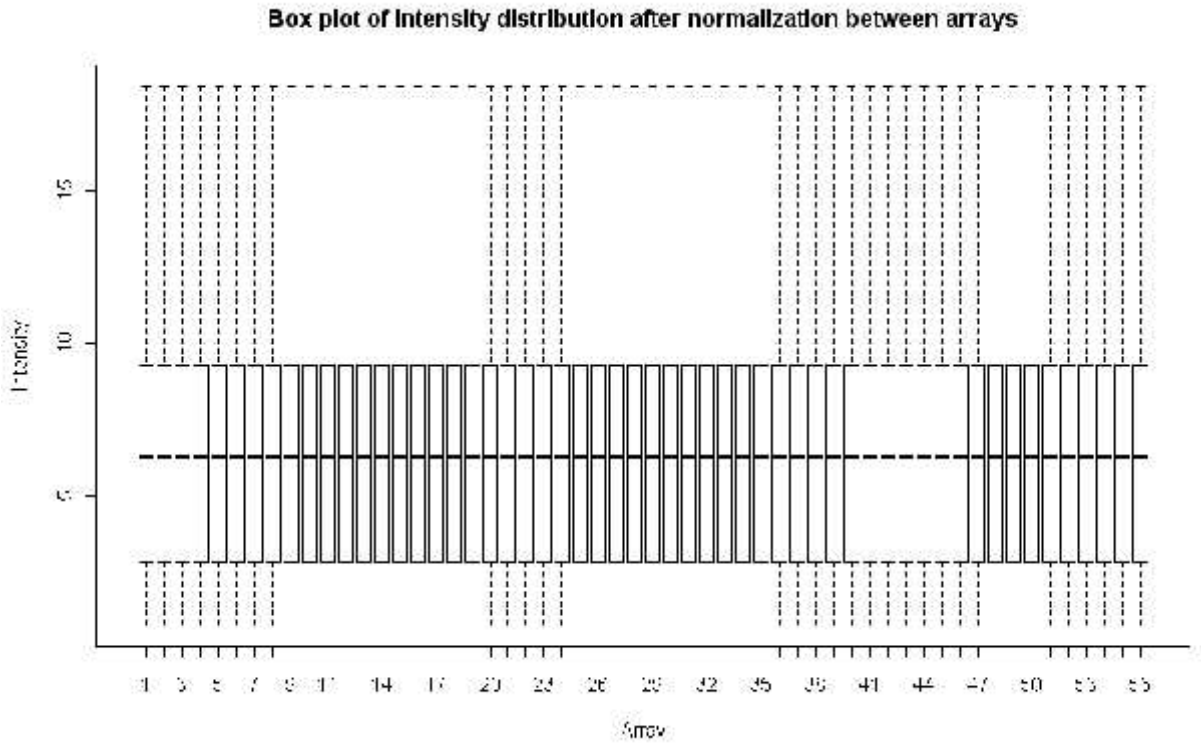

Figure 1: Box-plots showing distribution of expression intensities of microarray spots between arrays before normalisation (above) and after normalisation (below).

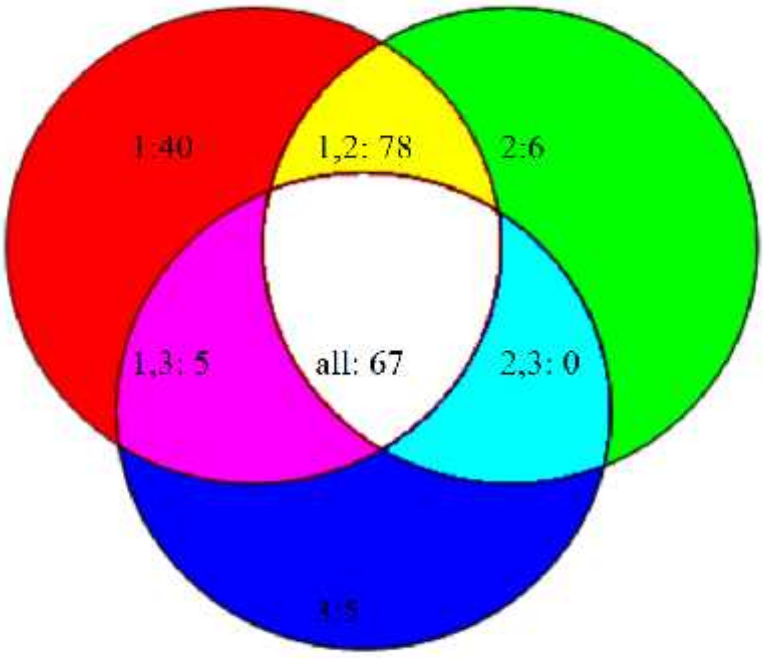

Figure 2: Venn diagram representing determined DE genes applying three different normalisation methods; (1) quantile normalization, (2) scale normalization and (3) variance stabilizing normalization.

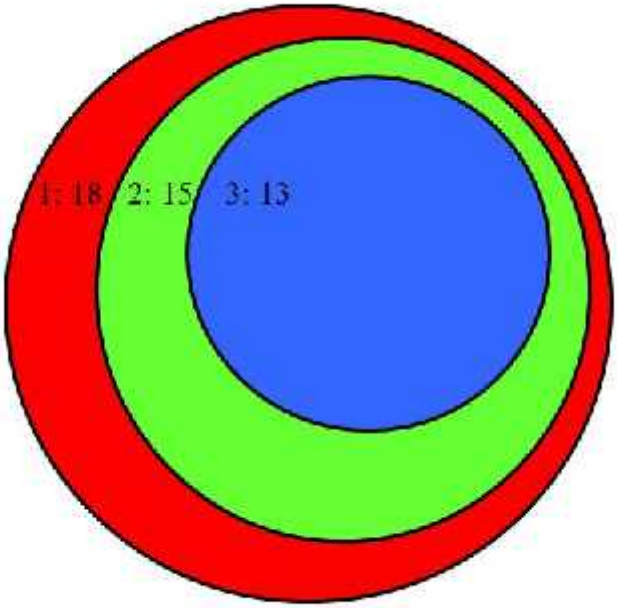

Figure 3: Venn diagram representing determined BINs of DE genes with three different normalisation methods; (1) quantile normalization, (2) scale normalization method and (3) variance stabilizing normalization.

Table 1: List of number of DE genes and BINs determined with three different normalisation methods; (1) quantile normalization, (2) scale normalization method and (3) variance stabilizing normalization.

No.	BIN	Biological function	Normalization method		
			1	2	3
1	1	photosynthesis	45	42	16
2	7	oxidative pentose phosphate pathway	1	0	0
3	9	mitochondrial electron transport, ATP synthesis	1	0	1
4	10	cell wall	5	3	1
5	11	lipid metabolism	5	4	1
6	13	amino acid metabolism	4	1	2
7	16	secondary metabolism	12	9	4
8	17	hormone metabolism	5	3	3
9	19	tetrapyrrole synthesis	6	7	5
10	20	stress	2	1	0
11	21	redox	3	2	0
12	23	nucleotide metabolism	1	0	0
13	26	misc	16	14	6
14	27	RNA	7	6	4
15	29	protein	8	6	3
16	33	development	3	2	0
17	34	transport	4	3	3
18	35	not assigned	63	48	28

4 CONCLUSIONS

Selected method for normalization of data between arrays can have substantial impact on the outcome of the analysis therefore optimal normalisation method should be carefully selected for the given experiment. For our experimental design quantile method was the most suitable, as it revealed the highest number of DE genes and BINs. On the other hand, the variance stabilizing normalization was not suitable for our data as vsn algorithm performs not only normalization but simultaneously also a background correction. This resulted in underestimation of DE genes as our data had a previous automatic step of background correction.

5 ACKNOWLEDGMENTS

This work was financially supported by the Slovenian Research Agency (ARRS), grant no. L4-2400-0401 and the Ministry of agriculture, forestry and food of Republic of Slovenia (MKGP).

6 REFERENCES

Bolstad, B.M., Irizarry R.A., Astrand, M., Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics, 19: 185-193.

Dolničar, P., Mavrič Pleško, I., Meglič, V. 2011. Long-term cold storage suppress the development of tuber necrosis caused by PVY-NTN. American journal of potato research, 4, 88: 318-323.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18 Supplement 1: S96-S104.

Smyth, G.K., Speed, T.P. 2003. Normalization of cDNA microarray data. Methods**,** 31: 265-273.

Yang, Y.H., Dudoit, S., Luu, P., Speed, T.P. 2001. Normalization for cDNA microarray data. In: Bittner, M.L., Chen, Y., Dorsel, A.N., Dougherty E.R. (eds.). Microarrays: Optical Technologies and Informatics, Proceedings of SPIE, 4266: 141-152.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 30,4: e15.

Yang, Y.H., Thorne, N.P. 2003. Normalization for two-color cDNA microarray data. In: Goldstein D.R. (ed.). Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes - Monograph Series, 40: 403-418.