Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 47 (3) 623-630 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Affymetrix GeneChip of Arabidopsis Thaliana shows less effects of G-stack probes

F. N. MEMON[++], Z. U. A. KHUHRO, A. P. HARRISON*

Institute of Mathematics and Computer Science, University of Sindh, Jamshoro

Received 15th March 2015 and Revised 27th June 2015

Abstract*:* This paper discusses the effects of probes with runs of 4 Guanines (a G-stack) on data taken from the GeneChip of a very common and widely used model organism in plant biology.We have examined the chip on probe level data and further investigated its normalised data. It is demonstrated that unlike variousGeneChips of mammalia, data from Arabidopsis Thaliana GeneChip, the ATH1-121501, is unaffected.

Keywords**:** Arabidopsis Thaliana, G-stack probes, Affymetrix, GeneChip

1. INTRODUCTION

The microarray technology, Affymetrix GeneChip for instance, has changed the scenario of gene expression profiling. It is a powerful functional genomic technology which enables the expressions of a large number of genes to be examined simultaneously in a single experiment. Memon (2010a) and Stalteri (2007) have discussed this technology in detail. The popularity of Affymetrix GeneChips is reflected by a large number of scientific papers based on this technology. Affymetrix GeneChips are not only adopted by the academic and research institutes but they are also popular among the pharmaceutical, biotechnological and diagnostic companies of the world.

Although Affymetrix has made a good reputation in different industries and its GeneChips are one of the popular microarrays, problems are found in them (Wu 2007, Upton 2008, Upton 2009, Langdon 2009, Memon 2010a, Memon, 2010b, Shanahan 2012). An improvement in future designs of GeneChips will be beneficial for all the industries and academic/non-academic research institutes/groups that use GeneChips.

Wu *et al.* (2007) have found some abnormal behavior on the surface of GeneChips because of G-stack probes. This was confirmed by Upton *et al.* (2008) who demonstrated that G-stack probes are not correlated with their member probes; however, they are correlated with each other regardless of their probe sets. Upton *et al.* (2008) and Memon *et al.* (2010a, 2010b) have examined the effects of G-stack probes at probe level data using various chip designs of different mammalians. However, Shanahan *et al.* have demonstrated that normalized data is also affected by G-stack probes which have been tested for one of the mammalian GeneChip, the HG_U133A (Shanahan 2012).

The Affymetrix GeneChips are available for a wide range of organisms that also includes a number of chips for various plants. As various mammalian GeneChip data is found to be affected by the G-stack probes, it is expected that the GeneChip data of plants may also be biased by G-stack probes. In this paper, the GeneChip of a widely used model organism in plant biology (Arabidopsis Thaliana) has been analyzed for the effects that are expected to be caused by the G-stack probes.

2. MATERIAL

The data for GeneChip of Arabidopsis Thaliana (i.e. ATH1-121501) have been selected for examination because it is considered as a model organism in plant biology. **(Table 1)** shows the statistics of the G-stack probes along with the number of affected probe sets and other information of the selected GeneChip. An affected probe set is the one that have at least one G-stack probe.

Table 1: The table shows the information about the chip design used in this study. The number of annotated probes and the number of G-stack probes are also listed which include both the main and control probes.

Organism	Arabidopsis Thaliana
Chip Design	ATH1-121501
Chip Size	712 x 712
Total Number of Annotated Probes	251,078
Total Number of G-stack Probes	5,839
Total Number of Probe Sets	22,810
Total Number of Affected Probe Sets	4,682
Release Date	July 2002

[++]Corresponding author: F. N. Memon, Email: farhatnm@usindh.edu.pk
*Department of Mathematical Sciences, University of Essex, UK

Data from 352 CEL files are used for analysis. These CEL files are chosen randomly from a number of GSEs that all belong to ATH1-121501 chip design.

3. METHOD

3.1 Association of G-stack probes with one another

This section presents the association of G-stack probes with each other regardless of their probe sets. This association is examined by obtaining the correlation coefficients among the G-stack probes and then generating a contour plot that shows the overall correlation surface for the selected chip design. The pipeline to obtain contour plots of the plants is the same that was used to analyze mammalian data sets in our previous work (Memon 2010b).

3.2 Level of hybridization of G-stack probes amongst the other member probes

The hybridization level of each of the G-stack probes is compared to the other members of their probe sets. Only those probe sets are considered here in which there are eleven probes in total, of which only one probe contains a single stack of exactly four guanines. The ranks of G-stack probes have been obtained against their member probes in each of the CEL files involved. For example, rank 11 is assigned to a G-stack probe if its intensity level is largest among the intensity values of all the eleven probes in its probe set. However, to present concise results, a median of ranks for each probe is taken over all the CEL files. Finally, a histogram is produced to visualize the frequencies of the ranks of G-stack probes intensities.

3.3 Analysis of summarized data

The summarized data has also been tested to verify the impact of G-stack on probe level data. The method used here is similar to the method presented in our previous work (Shanahan 2012) that measures the three parameters: expression values of the affected probe sets, differential expressions and the correlations between the expression values of the affected probe sets.

4. RESULTS AND DISCUSSION

4.1 Association of G-stack probes with one another

Fig. 1 shows the contour plot of correlation among the G-stack probes that have a G-stack on a particular position.

Table 1 illustrates that about 2.3% (5839/251,078) of the annotated probes contain a G-stack in Arabidopsis Thaliana array, affecting approximately 21% of the genes/probe sets. Although the ATH1-121501 chip design shows somehow similar fraction of G-stack probes as found in various mammalian chip designs (Memon 2010b), the correlation surface of the ATH1-121501 chip is entirely different. The plot in **Fig. 1** is showing poor correlations among the G-stack probes in ATH1-121501. The highest correlation value appeared in this plot is 0.34.

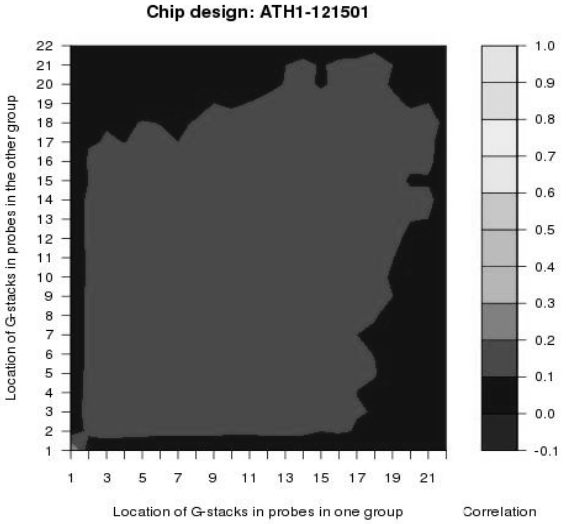

Fig. 1: Contour plot illustrating the change in average correlation coefficient values according to the position of the G-stack (with four Gs only) for a group of probes, in Arabidopsis Thaliana.

Unlike various GeneChips of mammalian, the contour plot of Arabidopsis Thaliana does not show much effect of G-stack probes at probe level data. Thus, a further investigation has been carried out on probe level data to verify the impact of G-stack probes on ATH1-121501 chip design.

4.2 Level of hybridization of G-stack probes amongst the other member probes

Fig. 2 shows the level of hybridization of G-stack probes in GeneChip of Arabidopsis Thaliana.

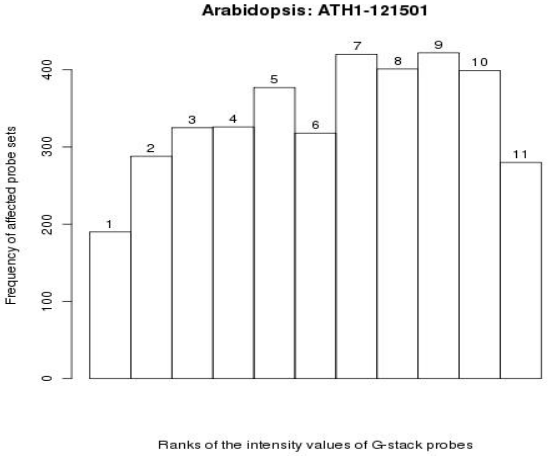

Fig. 2: The histogram shows the frequency of probe sets in which G-stack probe is at a particular rank in Arabidopsis Thaliana GeneChip, ATH1-121501.

Arabidopsis data shows an average level of intensity values of G-stack probes. Due to the poor correlation among the G-stack probe in Arabidopsis Thaliana (Figure 1), the level of hybridization was not expected to be high. Again, the probe level data is found to be unaffected in the ATH1-121501 chip.

The next section will demonstrate the effects of G-stack probes on summarized data to check if the summarized data is also unaffected for this particular chip.

4.3 Analysis of summarized data of ATH1-121501

Arabidopsis chip consistently showing less effects of G-stack probes on probe level data. This has been done by (i) examining the correlation surface (contour plot in Figure 1) and (ii) investigating level of hybridization of G-stack probes (Figure 2). As a final check, the summarized values of this chip have also been tested. The ATH1-121501 chip contains nearly 6,000 G-stack probes affecting about 21% of the annotated genes/probe sets. The total number of annotated genes on this chip is 22810. In addition to **(Table 1,a)** further statistics of affected probe sets according to the number of G-stack probes in them is given in **(Table 2).**

Table 2: A table listing the number of probe sets that have a specific number of G-stack probes.

No. of G-stack probes per probe set	0	1	2	3 or more
No. of probe sets	18,128	3749	776	157

For ATH1-121501, data for 162 GSEs are available at the local data repository that was created in late 2007. According to Upton *et al.* (2008), the most significantly affected experiments are those in which G-stack probes are most correlated with each other. Therefore, the average correlation among the G-stack probes has been calculated for 162 GSEs of ATH1-121501. The largest value, i.e. 0.4, is found for GSE5685 which is selected for this experiment.

The GSE5685 consists of 32 CEL files which can be grouped according to four different treatment protocols (Table 3). These protocols are considered as conditions to obtain differential expressions.

Initially, all the CEL files were used to obtain set of expression values, differential expressions and the correlation values among the expression values of the affected probe sets. All these parameters were measured for all possible combinations of a pair of treatment protocols. However, the results are presented only for one pair of protocols that consists of Avirulent pathogen and No Treatment. This selection is due to the largest variations that are observed between the protocols (i) Avirulent pathogen and (ii) No Treatment. Thus all the parameters are obtained using the CEL files of only these two protocols.

Table 3: A table listing the sample size according to the treatment protocols.

	Treatment Protocol	Sample size
1	Avirulent pathogen (Pseudomonas syringae ES4326 avrRpt2) infection	10
2	Mock treatment (10mM MgCl2 solution)	10
3	Virulent pathogen (Pseudomonas syringae ES4326) infection	10
4	No Treatment	02

4.3.1 Expression values

Fig. 3 illustrates the difference in RMA summarized values (Irizarry 2003) for Group A and Group B. Group A consists of probe sets that all have only one G-stack probe in them and these G-stack probes are included and excluded to get two sets of expression values to be analyzed. Group B consists of probe sets that all have no G-stack probe in them of which one randomly selected probe is included and excluded to get two sets of expression values to be analyzed. The number of probe sets in Group B is same as that in Group A.

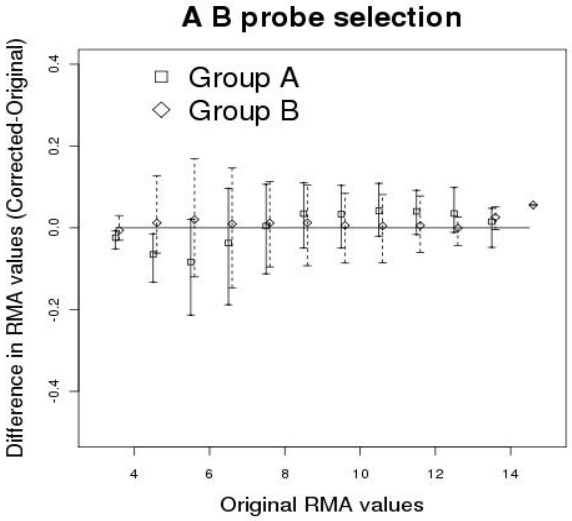

Fig. 3: The change in expression levels for Group A and Group B in dataset GSE5685.

Similarly, **(Fig. 4)** presents the difference in RMA summarized values for Group A2 and Group B2. Group A2 consists of probe sets that all have exactly two G-stack probes in them and these G-stack probes are included and excluded to get two sets of expression values to be analyzed and Group B2 consists of probe sets that all have no G-stack probe in them of which two randomly selected probes are included and excluded to get two sets of expression values to be analyzed. The number of probe sets in Group B2 is same as that in Group A2.

Removal of normal probes makes almost no difference on summarized values. Hence, the difference in expression values of Group B and Group B2 of ATH1-121501 is close to zero. On the other hand, some differences in expression values can be seen after removal of G-stack probes.

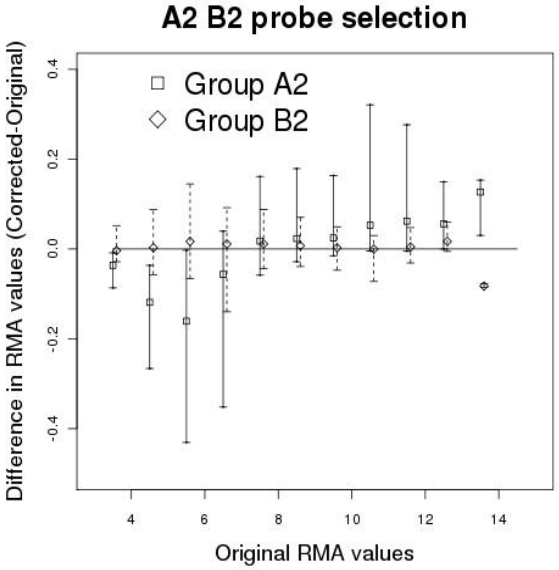

Fig.4: The change in expression levels for Group A2 and Group B2 in dataset GSE5685.

4.3.2 Differential expression

The effects of G-stack probes on fold change values are presented in **Fig. 5 to 12.** Figures 5 to 8 are showing the differences in fold change values for the probe sets with a particular number of G-stack probes.

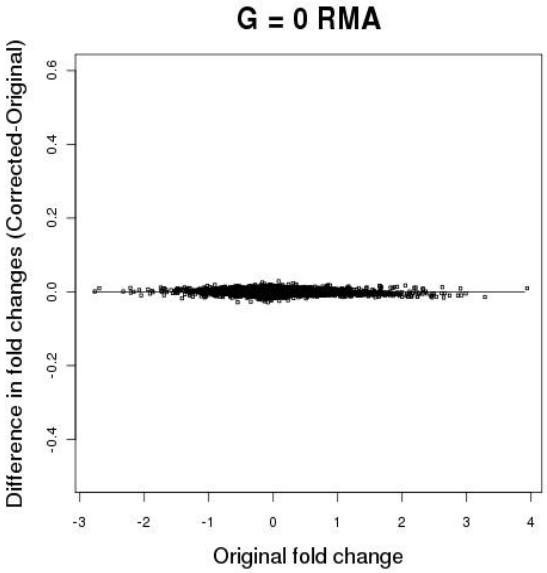

Fig. 5: The difference in fold change values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the fold change values of only those probe sets that have no G-stack probes in them.

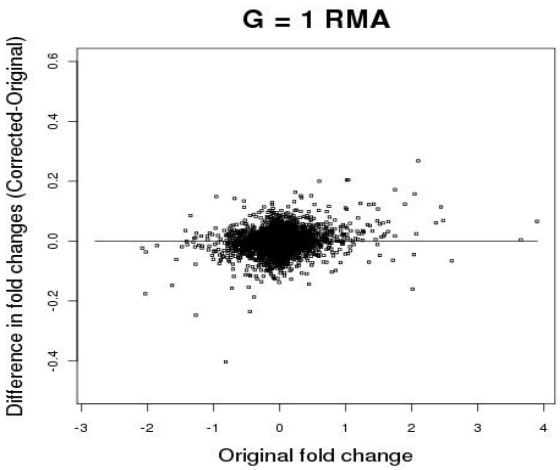

Fig. 6: The difference in fold change values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the fold change values of only those probe sets that have exactly 1 G-stack probe in them.

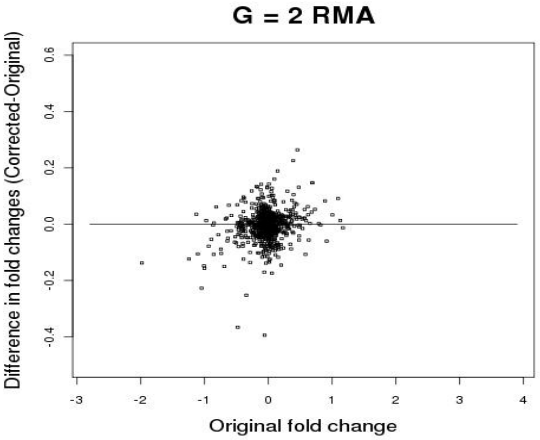

Fig. 7: The difference in fold change values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the fold change values of only those probe sets that have exactly 2 G-stack probes in them.

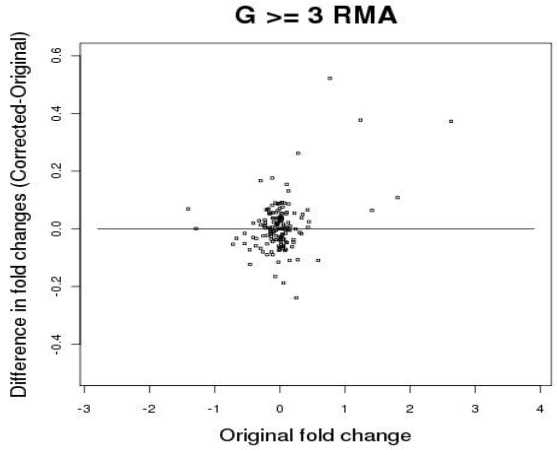

Fig. 8: The difference in fold change values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the fold change values of only those probe sets that have exactly 3 or more G-stack probes in them.

The difference in fold change values among Group A and Group B are illustrated in (**Fig. 9 and 10)** while that among Group A2 and Group B2 are presented in (**Fig.11 and 12).**

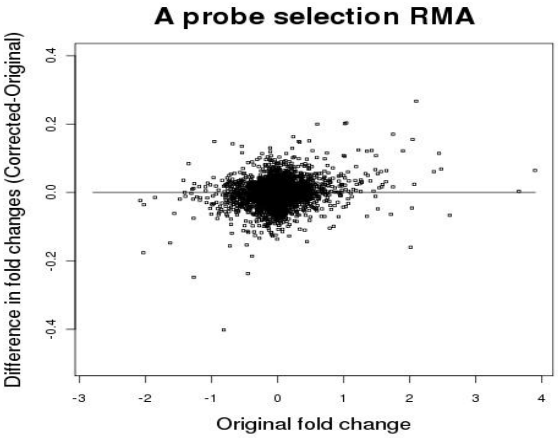

Fig. 9: The difference of the fold change values of probe sets in Group A data in dataset GSE5685.

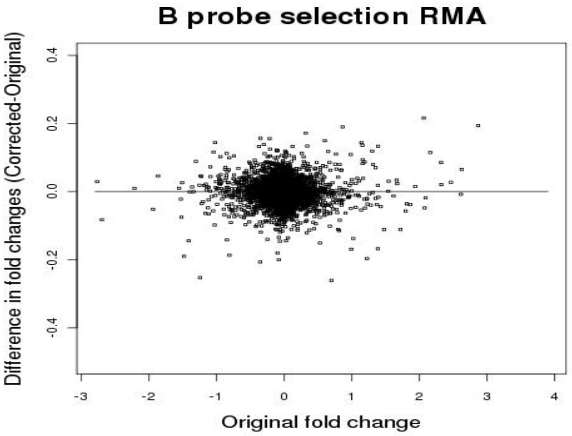

Fig. 10: The difference of the fold change values of probe sets in Group B data in dataset GSE5685.

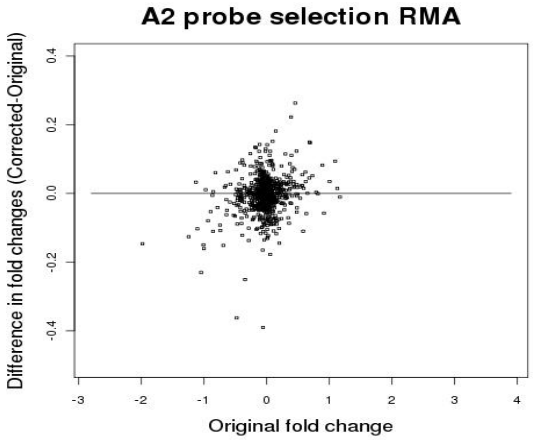

Fig. 11: The difference of the fold change values of probe sets in Group A2 data in dataset GSE5685.

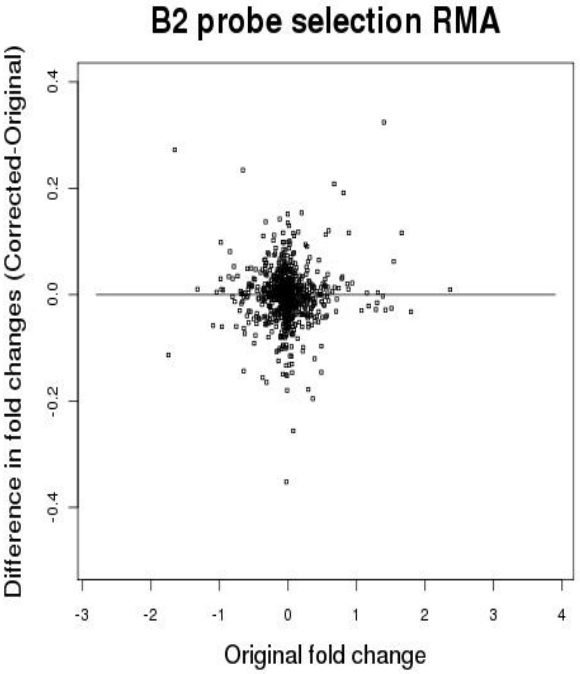

Fig. 12: The difference of the fold change values of probe sets in Group B2 data in dataset GSE5685.

4.3.3 Correlations among expression values of affected probe sets

The effects of G-stack probes on correlation values among the expression values of affected probe sets are presented in **(Fig. 13 to 18)**. Figures 13 to 16 are showing the difference in correlation values for the probe sets with a particular number of G-stack probes.

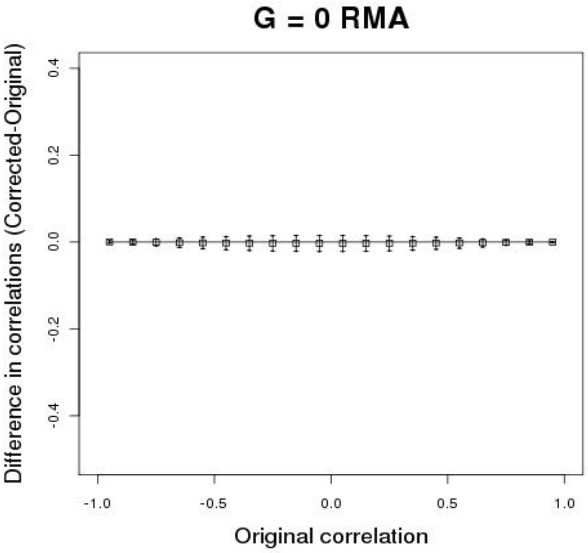

Fig. 13: The change in correlations among the expression values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the correlations among only those probe sets that have no G-stack probes in them.

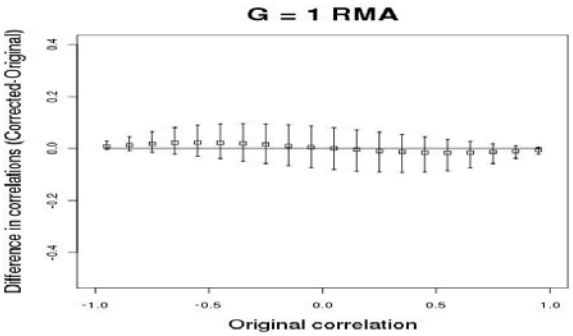

Fig. 14: The change in correlations among the expression values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the correlations among only those probe sets that have exactly 1G-stack probe in them.

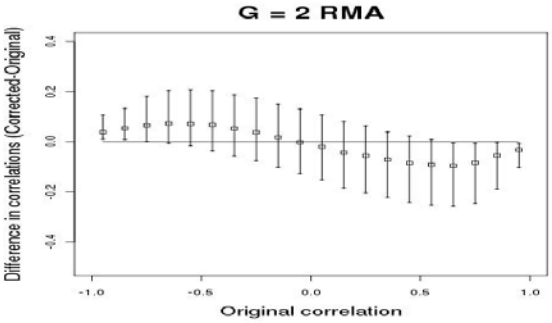

Fig. 15: The change in correlations among the expression values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the correlations among only those probe sets that have exactly 2 G-stack probes in them.

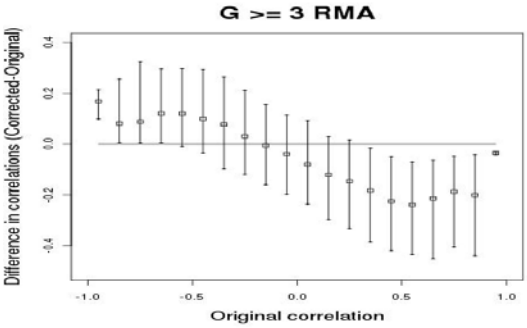

Fig. 16: The change in correlations among the expression values of probe sets before and after removal of G-stack probes in dataset GSE5685. The plot shows the correlations among only those probe sets that have exactly 3 or more G-stack probes in them.

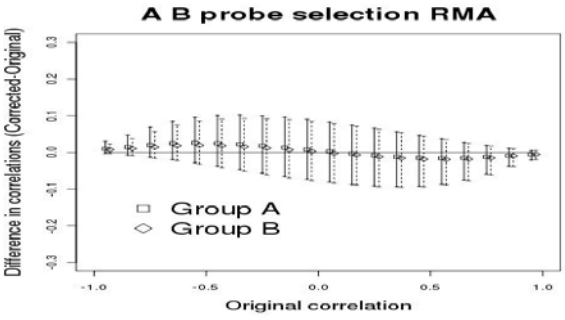

Fig.17: The change in correlation values for Group A and Group B data in dataset GSE5685.

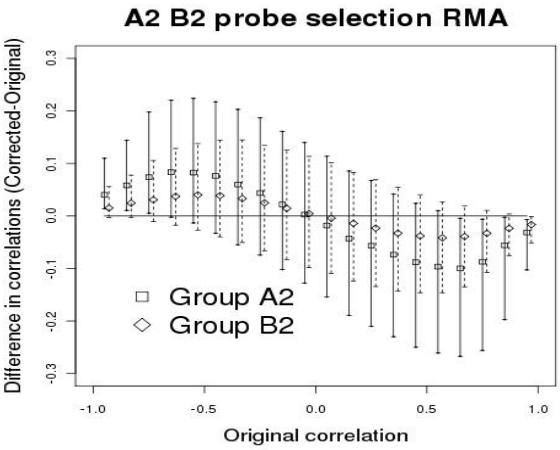

Fig. 18: The change in expression levels for Group A2 and Group B2 data in dataset GSE5685.

Similarly, the difference in correlation values among Group A & Group B and Group A2 & Group B2 are illustrated in Fig. 17 and Fig. 18 respectively.

All the plots **(from Fig. 3 to Fig. 18)** are showing the effects of G-stack probes on summarized data. The effects are considerably smaller than the effects on mammals chips (Memon 2010b). This is due to the facts that:

1. Shanahan *et al.* (2012) showed that the smallest effects on summarized data are seen in a GSE with an average G-stack probe-probe correlation of 0.41. As the G-stack probe-probe correlation of GSE5685 is 0.4 (a less affected GSE by the G-stack probes), it is natural to see smaller effects on parameters measured.
2. For human GeneChip, the HG_U133A, it is observed that slightly over one third probe sets having more than one G-stack probes are directly affected (Shanahan, 2012) and table 2 shows that about 25% probe sets in ATH1-121501 have at least one G-stack probe in them.

4.3.4 Different summarization methods on ATH1-121501 GeneChip data

The effects of G-stack probes on summarized data of human chip, the HG_U133A, have been presented in Shanahan, 2012. Initially it was found that the RMA method is biased by the G-stack probes and further verified that these variations do not appear only for RMA; other summarization methods are also affected. The four summarization methods, gcRMA (Naef 2003), MAS5 (Hubbell 2002), FARMS (Hochreiter 2006), and PLIER (Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, Affymetrix Technical report, Santa Clara 2005) were also tested and it was found that gcRMA and MAS5 are showing similar variations to that seen with RMA. However, FARMS and PLIER are found to be less affected.

The analysis for ATH1-121501 is also repeated for gcRMA, MAS5, FARMS, and PLIER to check if the performance of PLIER and FARMS is also better than the other methods. On ATH1-121501 data set, gcRMA and MAS5 are showing better performance and are less biased particularly for correlations when similar number of G-stack or normal probes is removed. After that PLIER is showing fewer changes.

It looks like all the summarized methods are biased due to G-stack probes. However, a specific method may give better performance on a particular dataset.

5 CONCLUSION

G-stack probes affect the probe level data as well as summarized data that has been examined through various animal GeneChip data in general and mammalian GeneChips in particular. This paper has focused on GeneChip of a widely used model organism in plant biology, i.e. Arabidopsis Thaliana, to identify if it is also affected by the G-stack probes.

The Arabidopsis Thaliana GeneChip (ATH1-121501) is showing poor correlation among these probes. The level of hybridization of G-stack probes is also found to be average. This shows that the probe level data is not affected by G-stack probes in ATH1-121501.

The ATH1-121501 data is then tested again for the effects of G-stack probes on summarized data. The effects of G-stack probes are examined on expression values, differential expression, and the correlations among the expression values of affected probe sets. Again the summarized data of ATH1-121501 is found unaffected. Although, the differences on the three parameters for ATH1-121501 are not very clear, small changes can be seen when G-stack probes are masked before summarization as compared to the removal of normal probes. These smaller differences as compared to that seen for mammals, are due to the fact that the individual GSEs of the ATH1-121501 are less affected by G-stack probes.

Different summarization methods that include RMA, gcRMA, MAS5, FARMS, and PLIER are tested for variations caused by the G-stack probes. It is found that all these methods are showing variations; however different methods are giving better performance on different data sets. Thus, all these methods are not reliable as their performance is dependent on the data set being analyzed.

5. ACKNOWLEDGEMENT

The authors are thankful for Charles Wallace Pakistan Trust for their support during this work.

REFERENCES:

Hochreiter S., D. A. Clevert, K. Obermayer (2006) *A new summarization method for Affymetrix probe level data*, Bioinformatics, 22, 943–949.

Hubbell E., W. M. Liu, R. Mei (2002) *Robust estimators for expression analysis*, Bioinformatics, 18, 1585–1592.

Irizarry R., B. Bolstad, F. Collin, L. Cope, B. Hobbs, T. Speed (2003) *Summaries of Affymetrix GeneChip probe level data*, Nucleic Acids Research, 31(4), 1-8, e15, doi: 10.1093/nar/gng015.

Langdon W., G. Upton, A. Harrison (2009) *Probes containing runs of guanines provide insights into the biophysics and bioinformatics of affymetrix GeneChips*, Briefings in bioinformatics, 10, 259-277.

Memon F. N., O. Sanchez-Graillet, G. J. G. Upton, A. M. Owen, A. P. Harrison (2010a) *Identifying the impact of G-Quadruplexes on Affymetrix 3' Arrays using Cloud Computing*, Journal of Integrative Bioinformatics , 7(2), 111, pp~1-9, doi:10.2390/biecoll-jib-2010-11.

Memon F. N., G. J. G. Upton, A. P. Harrison (2010b) *A comparative study of the impact of G-stack probes on various Affymetrix GeneChips of mammalian*, Journal of Nucleic Acids special issue on G-Quadruplex Nucleic Acids. Article ID 489736, 1-6, doi:10.4061/2010/489736.

Naef, F. and M. O. Magnasco (2003) *Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays*. Physical Review, E 68, 011906.

Shanaham H., F. N. Memon, G. Upton, A. Harrison (2012) *Normalized Affymetrix expression data are biased by G-quadruplex formation*, Nucleic Acid Research, 40(8), 3307- 3315.

Stalteri M. and A. Harrison (2007) *Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips*, BMC Bioinformatics, 8:13, *doi: 10.1186/1471-2105-8-13*.

Upton G., W. Langdon, A. Harrison (2008) *G-spots cause incorrect expression measurement in Affymetrix microarrays*, BMC Genomics, 9: 613, doi:10.1186/ 1471-2164-9-613

Upton G. J. G., O. Sanchez-Graillet, J. Rowsell, J. M. Arteaga-Salas, N. S. Graham, M. A. Stalteri, F. N. Memon, S. T. May, A. P. Harrison (2009) *On the causes of outliers in Affymetrix GeneChip data*, 8(3), Briefings in Functional Genomics and Proteomics , 199-212.

Wu, C., H. Zhao, K. Baggerly, R. Carta, L. Zhang (2007) *Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays*, Bioinformatics, 23(19), 2566–2572, doi: 10.1093/bioinformatics/btm 271.