Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 50 (3D) 156-160 (2018)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Effects of T-Quadruplex on Affymetrix GeneChip® Data (A Data Analytical Approach)

S. A. A. SHAH[++]*, F. N. MEMON*, Z.U.A. KHUHRO*, A. R. ABBASI**

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

Received 10th June 2018 and Revised 15th September 2018

Abstract: The Karst Hoogsteen theory that bases can also pair up in many different ways opens up another door of research for scientists. It was observed in previous research that four runs of contiguous Guanines form unusual structures called G-Quadruplex structures and some abnormal behaviors are found on microarray data particularly on Affymetrix GeneChip® data. These behaviors were associated with the formation of G-quadruplex structures due to the presence of G-stack probes. These probes were not correlated with their other member probes while they were correlated with each other regardless of the genes/probe sets. Hence, G-stack probes were suggested unreliable for gene expression measurement. This left another question that if G-G can affect GeneChip® data then interaction among other bases like T-T, A-A and C-C may also be problematic for GeneChips®.

The main objective of this research is to analyze the effects of thymine-thymine binding at probe level data of Affymetrix GeneChip®. Data of three different GeneChips® designed for organisms of three different kingdoms are downloaded from NCBI GEO repository. The data of Human (Kingdom Animalia), Arabidopsis Thaliana (Kingdom Plantae) and Escherichia Coli (Kingdom Bacteria) chips are used for analysis. Correlation among the T-stack probes was calculated to verify if they are behaving like G-stack probes or not. Our results suggest that thymine-thymine binding does not affect any of the chips taking into consideration. Hence it is all fine if T-stack probes are present on any GeneChip®. The scope of this research is to minimize the risk factor of GeneChips® that are being used in diagnostics and other purposes and communicate the effects of thymine-thymine with the scientists of biological world.

Keywords: Microarray, Affymetrix, GeneChip®, T-Quadruplex, thymine-thymine interaction.

1. INTRODUCTION

Microarray is said to be a powerful tool for measuring expression level of thousands of genes simultaneously (Kathleen, 2016).. Now, it has become universal and being adopted by the biologist worldwide (David *et al.,* 2006) A GeneChip® is a microarray manufactured by Affymetrix. The Affymetrix GeneChip® contains oligonucleotides called probes which recognize and hybridize with mRNA taken from target sample (Eun-Young *et al.,* 2016). GeneChip® is basically used for quantitative measurement of gene expression (Rafael and Bolstad1, 2003). On Affymetrix GeneChip®, a probe set consists of 11-20 probes, each of 25 bases long, to represent a gene (Laurent, 2004).

A **gene** is a region of DNA which is made up of nucleotides and is the molecular unit of heredity (Noble, 2008). The binding of Adenine (A) with Thymine (T) and Guanine (G) with Cytosine (C) has now become longstanding concept. Karst Hoogsteen also evaluated that the bases can also pair up in many different ways such as Guanine can bind with another Guanine and form unusual G4 structure called G-Quadruplex Structure (Farhat *et al.,* 2010). Currently, scientists are working on such structures which are formed due to unusual bonding between nucleotides such as Guanine binds with Guanine and Thymine with Thymine, etc.

It is also witnessed that GeneChip® data is biased due to the presence of G-stack probe (probes with continuous guanines) which are expected to form G-quadruplex structures on the surface of GeneChip® (Farhat *et al.,* 2010) (Farhat *et al.,* (2010) (Upton, *et al.,* 2008). It was observed that G-stack probes were poorly correlated with other members of their probe sets while they were highly correlated with each other. Hence, it was identified that G-stack probes on Affymetrix GeneChips® do not measure gene expression correctly (Farhat *et al.,* 2010) (Farhat *et al.,* (2010) (Upton, *et al.,* 2008). and suggested that these G-stack probes should be avoided in future GeneChip® designs..

As G-stack probes can cause misleading results of GeneChip®, it is expected that T-stack probes (probes with continuous thymine) may also form T-quadruplex structures and consequently may cause incorrect gene expression measurements – a similar behavior of G-stack probes.

[++] Corresponding author: Syed Akbar Ali Shah email: syedakbars@hotmail.com
*Institute of Mathematics & Computer Science University of Sindh, Jamshoro,
** Department of Fresh Water Biology and Fisheries, University of Sindh, Jamshoro, Pakistan.

This paper therefore presents the analysis of behavior of thymine-thymine interaction on raw/probe level data of Affymetrix GeneChips®.

2. MATERIAL

GEO (Gene Expression Omnibus) is a repository at National Centre for Biotechnology Information (NCBI) where data generated by different microarray platforms is archived and freely available (Tanya and Edgar. 2006). Affymetrix is one of the platforms whose data is available at GEO and could be used for further analysis.

The data of thousands of GeneChip® experiments for various organisms is freely available in public domain in form of CEL files. These CEL files can be downloaded from NCBI GEO (Gene Expression Omnibus) repository for analysis. This study initially focused on to analyze the effect of thymine-thymine interaction on Affymetrix GeneChip® designed for Homo-Sapiens (Human). Seven different designs of Human GeneChips® were tested to find the frequency of probes with continuous thymines (T-stack probes). However, HG_U95C, was selected for further examination as it has maximum number of T-stack probes and hence it was expected that the possible effects can be most prominently seen in HG_U95C. After getting the results of HG_U95C, data of two other GeneChip® designs were tested to verify our initial results. These two chip designs were of Arabidopsis Thaliana (an organism from Plantae kingdom) and E.Coli (Bacteria kingdom).

From manufacturing to the use of a GeneChip® in an experiment, a number of data files are generated including probe sequence file, CDF file, CEL files, etc. For this study, only Probe Sequence Data (Probe sequence files) and Experimental Data (CEL files) are required.

2.1 Probe Sequence Data (Probe Sequence File): It contains information about all the probes which are synthesized on a particular design of GeneChip®. This information includes sequences of probes and their x and y position on GeneChip®. This file helps to identify the probes of interest that means the probes having continuous thymine in their sequences. There is only one probe sequence file for a particular GeneChip® design. These files were downloaded from Affymetrix's website.

2.2 Experimental Data (CEL files): During an experiment, the hybridization level of each probe is scanned through a scanner and represented as a numeric value called intensity value of that probe. Intensity values of all the probes are finally stored in an electronic file called a CEL file. Each CEL file has a unique name starting with 'GSM' followed by unique number. Whereas 'GSE' followed by a unique number represents a particular biological experiment that may have multiple CEL files/GSMs. These files were downloaded from NCBI-GEO website.

3. METHOD

A complete pipeline is designed to analyze the behavior of T-stack probes or thymine-thymine interactions on GeneChip® data. This pipeline is similar to the one used to analyze the behavior of G-stack probes or guanine-guanine interaction (Farhat *et al.,* 2010) (Upton, *et al.,* 2008). The steps of the pipeline are briefly given below:

- Download the probe sequence file of the selected chip design.
- Probe sequence file is examined to filter out the T-stack sequences for the selected GeneChip® design. This step helps to extract the x and y position of probes of interest onto the chip. Finally these positions will be used to extract the intensity values of T-stack probes from the CEL files.
- Download the CEL files of that particular chip from NCBI-GEO repository.
- It was reported that probes with G-runs were outliers in their probe sets as they were not correlated with their other member probes whereas these G-run probes were highly correlated with each other while they are members of different probe sets. The next step is therefore to find the correlation among the T-stack probes. These T-stack probes are divided into groups according to the position of T-run in the sequences. It will help to analyze if there is any particular effect of position of T-run (in case T-stack probes show some effects on GeneChip® data. This division will create 22 groups of T-stack probes. For example, group 1 contains those T-stack probes in which position of T-run is 1.
- After classifying T-stack probes into groups, the correlation among all possible pair of these groups of T-stack probes are calculated which form a correlation matrix of order 22 by 22.
- These correlations among all the possible pairs of groups are illustrated by a contour plot to present the overall correlation surface of the selected GeneChip®.

4. RESULTS AND DISCUSSION

As mentioned earlier, Homo Sapiens (Human) GeneChip data was considered for the initial investigation. The probe sequence files of seven different designs of Human GeneChips® were downloaded and each was tested to identify the frequency of T-stack probes. **(Table 1)** is showing the seven designs of Human GeneChips® and statistics of annotated probes and T-stack probes on each chip.

Table-1: Table shows the Chip Designs, Total number of annotated probes, Total number of probes with T-run and the percentage (%) of T-stack probes.

Chip Design	Total No. of annotated probes	Total No. of T-stack probes	% of T-stack probes
HG_U133A	247965	20498	8.27
HG_U133_Plus_2	604258	54224	8.97
HG_U95A	201807	15868	7.86
HG_U95B	201863	19197	9.51
HG_U95C	201867	19681	9.75
HG_U95D	201858	19112	9.47
HG_U95E	201863	17757	8.80

Table 1 is providing the statistics of annotated and T-stack probes of various human GeneChip® designs. It can be seen from Table 1 that HG_U95B, HG_U95C and HG_U95D all have about 10% probes with T-stack in their sequences. However, HG_U95C has the highest fraction of T-stack probes i.e. 19681 out of 201867 probes. HG_U95C is therefore selected for detailed investigation. Table 2 is showing further statistics of T-stack probes and the probe sets that have at least one of these T-stack probes in the selected chip design; the HG_U95C.

Table 2 is showing that 19681 T-stack probes belong to 1211 different probe sets that means about 10% of the total probe sets have at least one probe with continuous thymine. Hence, any kind of effect of thymine-thymine measurement of about 10% genes which will be considered incorrect.

Table-2: Table shows the information about the structure of selected GeneChip® design (HG-U95C) the include chip size, total number of probes and total number of T-stack probes, etc.

Kingdom	Animalia (Animal)
Organism	Homo Sapiens (Human)
Chip Design	HG-U95C
Chip size	640x640
Total number of Annotated Probes	201867
Total number of T-stack Probes	19681
Total number of Probe Sets	12646
Total number of Probe Sets having at least one T-stack probe	1211

Correlation among the groups of T-stack probes are calculated which are presented in Table 3 and the overall correlation surface is illustrated in Figure 1 that shows the effects of T-stack probes on HG_U95C data.

Table 3 is showing that the largest correlation value is 0.05 which is a diagonal value and is the average correlation of group 1 with group 1 that means the correlation among all the T-stack probes in which T-run is at position 1 is showing the highest correlation value. It is noted that position 1 represents to the free end of the probe so the highest correlation value might be the result of free end of T-stack probe that have more chances of interaction. Whereas, the smallest value is among the T-stack probes in which T-run is at position 22 which is the fixed end of the probe. The smallest correlation value is 0.0049. The entire correlation matrix (Table 3) is not showing significant effects of T-stack probes on probe level data of HG_U95C.

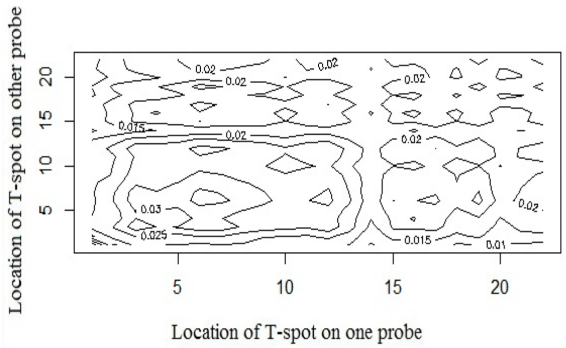

Fig.1: shows the contour map of average correlation of HG_U95C.

Although it is observed from **(Table 3) and (Fig.1)** that T-stack probes are poorly correlated with each other, other chip designs are tested further to verify if only HG_U95C (human chip design) is unaffected or other chips are also unaffected by thymine-thymine interactions. Therefore, two more chip designs were selected from different kingdoms: Plantae and Bacteria. These chips are particularly designed for Arabidopsis Thaliana and E. Coli. The same process is carried out on the data of the two selected chip designs in order to compare the results of three organisms of three different kingdoms.

The information about the structure of E. Coli and Arabidopsis chips are presented in **(Table 4 and 5)** respectively. Table 4 is showing that in E. Coli chip design, 6519 out of 112488 probes have T-stack in them that make about 6% of the annotated probes. Furthermore, 38% of the probe sets contain at least one T-stack probe in them.

Table-3: The table is illustrating the correlation matrix (order 22 x 22) where average correlation among all the possible pairs of groups of T-stack probes is presented for HG_U95C.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.02	0.02	0.02	0.02	0.02	0.01
2	0.04	0.03	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
3	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.04	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
4	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
5	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.01
6	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.01
7	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.01
8	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.01
9	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.01
10	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
11	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01
12	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.01
13	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
14	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.01
15	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01
16	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.01
17	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
18	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
19	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
20	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
21	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
22	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0

Table-4: Table shows the information about the E.Coli GeneChip®.

Kingdom	Bacteria
Organism	E.Coli
Chip Design	Genome 2.0 Array.
Chip size	478x478
Total number of Annotated Probes	112488
Total number of T-stack Probes	6519
Total number of Probe Sets	10208
Total number of Affected Probe Sets having at least one T-stack probe	3832

Similarly, **(Table 5)** is showing that Arabidopsis Thaliana chip design have 17191 out of 251078 probes with T-run in their probe sequences. Therefore, in Arabidopsis Thaliana chip about 7% probes are considered as T-stack probes which are members of about 50% probe sets.

Table-5: Table shows the information about the Arabidopsis Thaliana GeneChip®.

Kingdom	Plantae (Plant)
Organism	Arabidopsis Thaliana
Chip Design	ATH1-121501
Chip size	712x712
Total number of Annotated Probes	251078
Total number of T-stack Probes	17191
Total number of Probe Sets	22810
Total number of Affected Probe Sets having at least one T-stack probe	11318

(Fig. 2 and 3) are showing overall correlation surfaces of E. Coli and Arabidopsis Thaliana chips respectively.

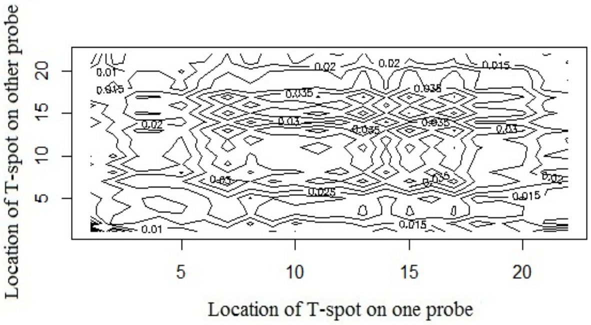

Fig. 2: shows the contour map of average correlation of E.Coli

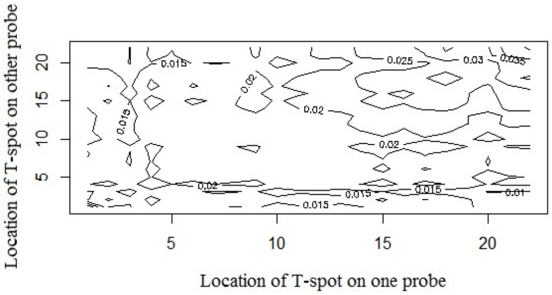

Fig.3: shows the contour map of average correlation of Arabidopsis

It can be seen from **(Fig. 1, 2, and 3)** that all the three chip designs that represent three different organisms as well as three different kingdoms are showing poor correlation among T-stack probes. Although Arabidopsis Thaliana GeneChip contains about 50% of genes/probe sets have atleast one T-stack probe but despite this large fraction, the chip data is showing poor correlation among the T-stack probes. Similar results are seen in other two chip designs with 38% and 10% probe sets with at least one T-stack probe. This similar behavior on different chip designs gives the impression that the Affymetrix GeneChips are generally unaffected by the presence of T-stack probes.

5. CONCLUSION

Besides the usual base pairing of cytosine - guanine and adenine – thymine, pairing up of same bases has become a constant issue for research scientists as the previous results of other research depicted. Keeping in view the same issue, this research is carried out to find the effects of thymine binding with another Thymine on three different GeneChip® data. Probe sequence data of all available Human GeneChip® was collected and passed through in-house tool.

HG-U95C was found with maximum number of T-stack probes and selected for further analysis. 504 CEL files (Experimental Data) of HG-U95C was downloaded from NCBI-GEO repository for further analysis.

In-house tools were developed for calculating correlation values among the T-stack probes regardless of their probe sets. To test if there is any particular effect of position of T-run in probe sequences, the T-stack probes were divided into groups according to the position of T-run.

Finally, poor correlation is observed among T-stack probes. Furthermore no particular effect is seen for the position of T-run within the probe sequences of T-stack probes. The similar experiment is repeated on chip designs of two other organisms (i.e., Arabidopsis Thaliana and E.Coli) from different kingdoms in order to verify the results of Human GeneChip® data. Hence, it is verified that in general GeneChip® data is unaffected by the presence of T-stack probes. This result minimizes the risk factor associated with GeneChip® and left satisfactory results for medical experts and biological scientists who were using this GeneChip® for various purposes.

REFERENCES:

David B. A., Xiangqin Cui, G. P. Page and M. Sabripour. (2006) *Microarray data analysis: from disarray to consolidation and consensus*. Nature Reviews Genetics 7, 55-65 | doi:10.1038/nrg1749.

Eun-Young Cha,1 Hye-Eun Jeong,1 Woo-Young Kim, 1Ho Jung Shin,1 Ho-Sook Kim,1,2 and Jae-Gook (2016) Shin1, 2 *Brief introduction to current pharmacogenomics research tools.* Transl Clin Pharmacol.; 24(1):13-21.

Farhat N. M., Anne M. Owen, O. Sanchez-Graillet, Graham J. G. Upton and A.Harrison. (2010) *Identifying the impact of G-Quadruplexes on Affymetrix 3' Arrays using Cloud Computing*. Journal of Integrative Bioinformatics, 7(2):111.

Farhat N. M., Graham J. G. Upton, and A. P. Harrison. (2010) *A Comparative Study of the Impact of G-Stack Probes on Various Affymetrix GeneChips of Mammalia*. SAGE-Hindawi Access to Research Journal of Nucleic Acids Volume, Article ID 489736, 6Pp doi:10.4061/2010/489736.

Kathleen M. (2016).Eyster. *DNA Microarray Analysis of Estrogen-Responsive Genes.* Volume 1366 of the series Methods in Molecular Biology 115-129

Laurent G., (2004) Leslie Cope2, Benjamin M. Bolstad3 and Rafael A. Irizarry4. *Affy—analysis of Affymetrix GeneChip data at the probe level* Vol. 20 no. 3, 307–315
doi: 10.1093/bioinformatics/btg405.

Noble, D. (2008) "Genes and causation, " *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1878, 3001–3015**.**

Rafael A. I., B. M. Bolstad1, (2003) Francois Collin2, Leslie M. Cope3 , Bridget Hobbs4 and Terence P. Speed4,5. *Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Research*, , Vol. 31, No. 4 15
doi: 10.1093/nar/gng015.

Tanya B and R Edgar. (2006) *Gene Expression Omnibus (GEO): Microarray data storage, submission, retrieval and analysis.* National Institute of Health Public Access Author, Methods Enzymol, 411:352-369.

Upton, G. J., W. B. Langdon, and A. P. Harrison, (2008) "G-spots cause incorrect expression measurement in Affymetrix microarrays," *BMC Genomics*, vol. 9, no. 1, p. 613.

Following data is added in conference proceeding paper as suggested in conference.

In proceeding paper, the result of T-stack probes was shown only on Affymetrix Human GeneChip® data.

It was suggested in conference that some more results of other chips should be added to verify the results.

In extended paper, data of two (02) more GeneChips® of other organisms of different kingdom (i.e *Arabidopsis thaliana* and *E.coli*) is added.