*Sindh Univ. Res. Jour. (Sci. Ser.) Vol.50 (001) 01-06 (2018)*

http://doi.org/10.26692/sujo/2018.01.0001

**SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)**

# Statistical analysis of Bacteria GeneChips for possible effects caused by Hoogesteen hydrogen bonding between Guanines

F. N. MEMON[++], Z. U. A. KHUHRO, Q.U.A. NIZAMANI, F. PANHWAR*, N. H. ARIJO**

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

Received 26th January 2017 and Revised 18th October 2017

**Abstract:** According to Watson-Crick base pairing, Guanine binds to Cytosine and Adenine binds to Thymine to form a double helix DNA molecule; however, it is observed that guanine can also bind with another guanine through hoogesteen hydrogen bonding at about 90 degrees. Thus a tetrad is formed if four guanines are attached with each other. These tetrads are the cause of the formation of unusual structures called G-quadruplex structures. These structures not only formed by a single guanine rich nucleic acid sequence but their formation is also possible by two or four parallel sequences collectively.

These structures (or guanine rich sequences through which they are formed) affect a well known technology named Affymetrix GeneChip which is popular for gene expression measurement effectively and cheaply. Since the nucleic acid sequences are placed very closely on GeneChip, they are able to interact with each other and then form G-quadruplex structures. In this case, they are not ready to hybridize with their target sequences. It is found that raw or hybridization level data of human, mouse and rat (mammalian in general) are affected by guanine rich sequences. This paper exhibits that GeneChips of two Bacterias are also affected by guanine-guanine interaction and hence measurement of gene expressions using these chips could be misleading.

**Keywords**- Bacteria, E.Coli, Pseudomonas Aeruginos., G-quadruplex structures,, G-stack probes, Affymetrix GeneChip.

## 1. INTRODUCTION

Two nucleotide bases that interact with each other through hydrogen bonding is known as base pairing. According to Watson-Crick base pairing, a guanine binds to a cytosine and an adenine binds to a thymine (Watson 1953). This Watson-Crick base pairing is important for the formation of double helix DNA structure (Marky 2010). However, it is also known that a guanine can bind to another guanine through a different type of bonding called Hoogsteen hydrogen bonding. Hence, in a nucleotide sequence with frequent occurrence of guanines, there is possibility of guanine-guanine interaction that can form unusual structures called G-quadruplex structures (Siddiqui-Jain 2002; Huppert 2005; Burge 2006; Qin 2008). These unusual structures can be formed in a single stranded nucleotide sequence; however, two or four parallel sequences with runs of guanines can also form these structures (Bates 2007).

In 2008, Upton *et. al.*, reported that an Affymetrix GeneChip (HG_U133A) is affected due to the presence of probes that have four or more continuous guanines in their sequences (they referred them G-spot probes). Upton *et. al.*, associated the misbehavior of G-spot probes with the possibility of the formation of G-quadruplex structure on the surface of GeneChip.

These effects were further tested on various GeneChip designs for human, mouse and rat and it was reported that the mammalian GeneChips are affected by the probes with runs of guanines (Memon 2010b). Memon et. al called them G-stack probes. It was further found that not only raw/probe level data but probe set/summerized data is also biased due to G-stack probes (Shanahan 2012).

A GeneChip is a microarray that is produced by the Affymetrix company and is commonly used for gene expression measurement. It is a quick and relatively inexpensive tool for gene expression measurement hence is popular among scientific community. However, Naef (2003), Wu (2007), Stalteri (2007), Upton (2008), Upton (2009), Langdon (2009), Memon (2010a, 2010b), Shanahan (2012), Fasold (2012), Fasold (2014), Memon (2015) have reported different problems. These problems, if rectify, can help in improvement of future chip designs.

A gene is represented on a GeneChip by a group of 25 bases long oligonucleotides (also called probes) in 2-dimensinal arrangement. Each group typically consists of 11 to 16 probes (Upton 2008) and is called a probe set. The size of a chip/array is about a thumb nail that contains millions of probes which are placed very

[++]Corresponding author: F. N. Memon, Email: farhatnm@usindh.edu.pk
* Department of Pharmacology, University of Sindh, Jamshoro
** Institute of Information and Communication Technology, University of Sindh, Jamshoro

closely. It is due to the close proximity of probes that if neighboring probes have continuous guanines in their sequences, possibly they are able to form G-quadruplex structure and that can cause incorrect gene expression measurement.

It is observed that G-stack probes are outliers in their probe sets (Wu 2007, Upton 2008, Upton 2009, Memon 2010a, Memon 2010b). Various GeneChip designs of human, mouse and rat are found affected due to these probes (Memon 2010b). They behave different as compared to other member probes in mammalian chip designs. The hybridization level of these problematic probes are found relatively higher than the other member probes. On the other hand, these probes are highly correlated with each other regardless of their probe sets.

This paper focuses on two GeneChips that were designed for two different bacterias: (i) E. Coli and (ii) Pseudomonas Aeruginos. The effects of G-stack probes have been measured for the selected GeneChips using their hybridization level data publicly available at NCBI GEO website.

## 2. MATERIAL

The GeneChip data of two different chips is taken for examination which are both designed for the organisms from the kingdom of Bacteria. These GeneChips are (i) E.Coli Genome 2.0 Array and (ii) Pseudomonas Aeruginosa Genome Array.

A number of different data files are normally generated for a GeneChip from manufacturing to its use in an experiment; however, only two types of data files were required for this study:

1. Probe sequences file (in FASTA format) that contains the 25 bases long sequences of a chip and their positions on the 2-D array.
2. CEL files that contain hybridization level of each probe during an experiment.

355 CEL files for each chip are used to examined the effects of G-stack probes. These CEL files are chosen randomly from 22 GSEs for P. Aeruginosa and from 25 GSEs for E.Coli. The selection of CEL files and GSEs is an entire random selection.

## 3. METHOD

A pipeline is designed for this study that involves a number of different steps which are described below. The complete pipeline is performed on the data of each chip separately.

### 3.1 Identification of problematic probes (the G-stack probes)

The probes having a single run of exactly four guanines are filtered out from the probe sequence file of the relevant GeneChip.

### 3.2 Extraction the positions of G-stack probes

This step is performed to get the x and y positions of G-stack probes on GeneChip so that the hybridization level of these probes can be extracted from the CEL files for examination.

### 3.3 Distribution of G-stack probes into group

Each G-stack probe is composed of 25 nucleotides and a run of exactly four guanines is present in this sequence. This G-run may be the first four nucleotides of the sequence that means the G-run is at position 1 in probe sequence. Similarly, the G-run may be at position 2, 3, 4, 22. Hence 22 groups are created and all the identified G-stack probes are placed in these groups according to the place of G-run in the probe sequences.

### 3.4 Association of G-stack probes with one another

It is obvious that a group of probes that represent a single gene on a GeneChip should behave in similar way and hence they must be correlated with each other. But literature has demonstrated that G-stack probes are not correlated with their member probes and their hybridization level is also showing some kind of biasness; they are relatively higher than other member probes. Furthermore, the G-stack probes are highly correlated with each other while they are members of different probe sets.

To examine the association of G-stack probes with each other, the average correlation coefficient is measured among all the possible pairs of two groups of G-stack probes using the CEL files. This correlation measurement is done among the groups that are created in step 3.3 and as a result of this step a 22x22 correlation matrix M is created. A contour plot is then generated to illustrate the overall correlation surface of the GeneChip using correlation matrix.

### 3.5 Effects of the position of G-run within the G-stack probes

It is also examined that if the position of G-run within the G-stack probes has any effect. This is done by analyzing the correlation matrix M. It is expected to get the high correlation values at the diagonal of M. These diagonal values represent the correlation among the two groups in which position of G-run is same in the G-stack probes' sequences of the two groups.

## 4. RESULTS AND DISCUSSION

As compared to the GeneChips of human, mouse, rat and Arabidopsis Thaliana, the two chips of bacteria are relatively small in size. **(Table 1)** shows the relevant information of the two chip designs. This information provides the information regarding the structure of the selected chips. It also shows the frequency of G-stack probes and the affected probe sets (i.e. probe sets contain at least one G-stack probe).

**Table 1: The table illustrates the statistics of probes and other information of the selected GeneChips.**

| Scientific Name of Organism | Escherichia coli | Pseudomonas Aeruginos |
|---|---|---|
| Chip Design | E.Coli Genome 2.0 Array | Pseudomonas Aeruginosa Genome Array |
| Chip Size | 478*478 | 403*403 |
| Total Number of Annotated Probes | 112,488 | 77,674 |
| Total Number of G-stack Probes | 318 (19 control probes are included) | 1830 (2 control probes are included) |
| % of Probes with G-stack | 0.28 | 2.36 |
| Total Number of Probe Sets | 10,208 (10 Control PS) | 5,900 (15 Control PS) |
| Total Number of Affected Probe Sets | 199 (9 Control PS) | 1,416 (2 Control PS) |
| % of Affected Probe Sets | 1.95 | 24 |

The frequencies of G-stack probes in the two selected chips are very different. The GeneChip of E.Coli has 0.28% (318/112,488) of annotated probes with run of guanines which affects about 2% of the total probe sets. On the other hand, the GeneChip of P. Aeruginos has a higher frequency of G-stack probes that is 1830 out of 77,674 annotated probes and it affects 24% of the total probe sets. The control probes with G-stacks (16 in E. Coli and 2 in P. Aeruginos) are not included in the examination. **(Fig. 1 and 2)** show the frequency of total G-stack probes according to the positions of stack of guanines in the nucleotide sequences of these G-stack probes.

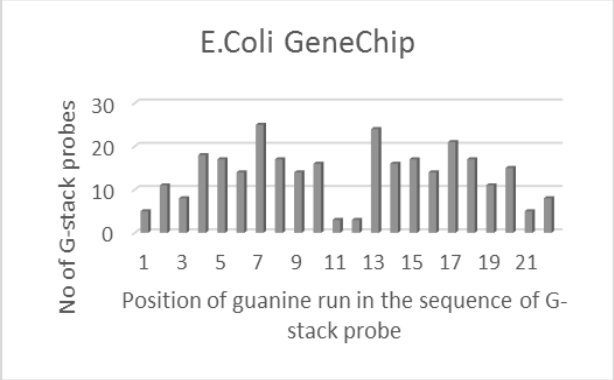


**Fig. 1: This plot is illustrating the frequency of G-stack probes according to the position of G-run in the nucleotide sequences, in E.Coli Genome 2.0 Array.**

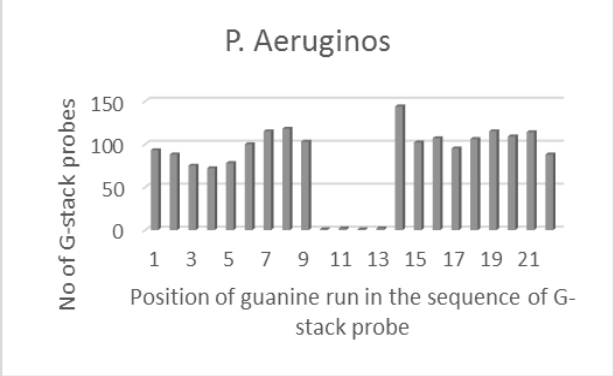


**Fig. 2: This plot is illustrating the frequency of G-stack probes according to the position of G-run in the nucleotide sequences, in Pseudomonas Aeruginosa Genome Array.**

Both the **(Fig. 1 and 2)** show a very small number of G-stack probes in which G-run is at middle positions of sequences. This was also found in new chip designs of mouse where no probe has G-stack at position 11 in MOE430A and no probe contains G-stack at position 11 and 12 in MOE430B.

Furthermore, **(Fig. 3 and 4)** are showing association of G-stack probes with one another in the two selected GeneChips. The association is represented by contour plots to illustrate the overall correlation surface of the chips. The correlation matrices of E.Coli and P.Aeruginos are also presented by **(Tables 2 and 3)** respectively.

**Table 2: Correlation matrix of order 22x22 that is showing the correlation among different groups of G-stack probes for Escherichia Coli Genome 2.0 array.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.5 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 |
| 2 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 |
| 3 | 0.5 | 0.4 | 0.6 | 0.6 | 0.6 | 0.5 | 0.5 | 0.5 | 0.6 | 0.5 | 0.7 | 0.6 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.2 | 0.2 | 0.2 | 0.2 |
| 4 | 0.5 | 0.4 | 0.6 | 0.5 | 0.6 | 0.5 | 0.4 | 0.5 | 0.6 | 0.5 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 |
| 5 | 0.6 | 0.4 | 0.6 | 0.6 | 0.6 | 0.6 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.6 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.2 | 0.3 | 0.2 |
| 6 | 0.5 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.5 | 0.6 | 0.5 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 7 | 0.4 | 0.4 | 0.5 | 0.4 | 0.5 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 |
| 8 | 0.5 | 0.4 | 0.5 | 0.5 | 0.5 | 0.5 | 0.4 | 0.5 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 |
| 9 | 0.5 | 0.4 | 0.6 | 0.6 | 0.6 | 0.6 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.6 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.2 | 0.3 | 0.2 |
| 10 | 0.5 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.6 | 0.4 | 0.4 | 0.5 | 0.4 | 0.4 | 0.3 | 0.3 | 0.2 | 0.3 | 0.2 |
| 11 | 0.6 | 0.5 | 0.7 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.7 | 0.7 | 0.8 | 0.7 | 0.5 | 0.6 | 0.6 | 0.5 | 0.5 | 0.5 | 0.4 | 0.3 | 0.3 | 0.3 |
| 12 | 0.5 | 0.4 | 0.6 | 0.5 | 0.6 | 0.5 | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.5 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.2 | 0.2 | 0.2 |
| 13 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 14 | 0.4 | 0.3 | 0.5 | 0.4 | 0.5 | 0.4 | 0.3 | 0.4 | 0.5 | 0.4 | 0.6 | 0.5 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 | 0.2 |
| 15 | 0.4 | 0.3 | 0.5 | 0.4 | 0.5 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.4 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.2 | 0.3 | 0.2 |
| 16 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 17 | 0.3 | 0.2 | 0.4 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | 0.5 | 0.4 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 |
| 18 | 0.3 | 0.2 | 0.4 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.4 | 0.3 | 0.5 | 0.4 | 0.3 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 19 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.2 | 0.2 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 | 0.3 | 0.2 | 0.3 | 0.2 |
| 20 | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 |
| 21 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.2 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.2 | 0.3 | 0.1 | 0.2 | 0.2 |
| 22 | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 |

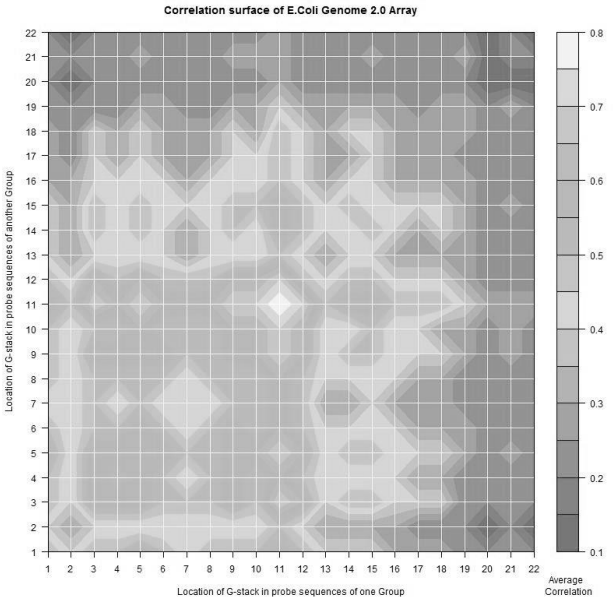


**Fig. 3: Contour plot illustrating the change in average correlation coefficient values according to the position of the G-stack (with four Gs only) for a group of probes, in E.Coli Genome 2.0 Array.**

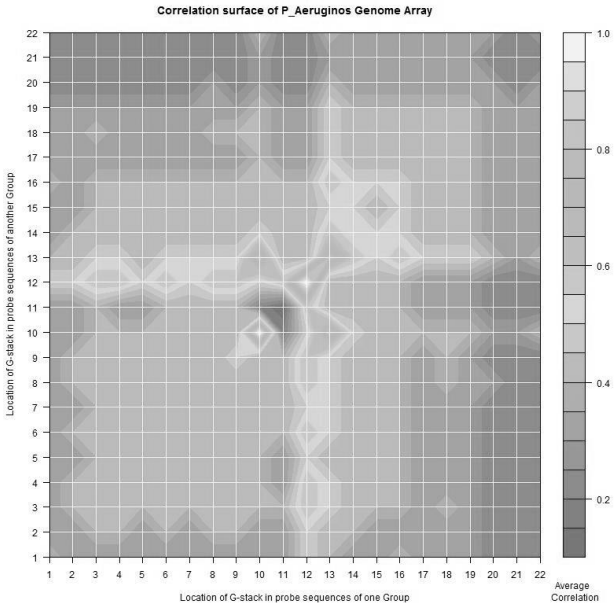


**Fig. 4: Contour plot illustrating the change in average correlation coefficient values according to the position of the G-stack (with four Gs only) for a group of probes, in Pseudomonas Aeruginosa Genome Array.**

**Table 3: Correlation matrix of order 22x22 that is showing the correlation among different groups of G-stack probes for Pseudomonas Aeruginos Genome array.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.5 | 0.4 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 2 | 0.3 | 0.3 | 0.4 | 0.4 | 0.3 | 0.4 | 0.3 | 0.4 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 | 0.2 |
| 4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 5 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 6 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 7 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 8 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 | 0.3 | 0.3 | 0.2 | 0.2 |
| 9 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.4 | 0.6 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 | 0.2 |
| 10 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 1 | 0.2 | 0.5 | 0.8 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.4 |
| 11 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.2 | 0.1 | 0.6 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 12 | 0.5 | 0.5 | 0.6 | 0.6 | 0.5 | 0.6 | 0.5 | 0.6 | 0.6 | 0.5 | 0.6 | 1 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 13 | 0.4 | 0.4 | 0.5 | 0.5 | 0.4 | 0.5 | 0.5 | 0.5 | 0.5 | 0.8 | 0.4 | 0.4 | 0.8 | 0.6 | 0.5 | 0.6 | 0.5 | 0.5 | 0.5 | 0.4 | 0.3 | 0.4 |
| 14 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.4 | 0.6 | 0.5 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 15 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.5 | 0.5 | 0.4 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 16 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.6 | 0.5 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 17 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 18 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 19 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.3 |
| 20 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.3 | 0.2 | 0.2 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 |
| 21 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.2 |
| 22 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.4 | 0.2 | 0.2 | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 |

Table 1 demonstrates that the E.Coli and P. Aeruginos GeneChips use a bigger feature size as compared to that of various mammalian chips (Memon 2010b) and Arabidopsis Thaliana chip (Memon 2015). As mammalian and Arabidopsis chips, the number of probes containing a stack of four guanines is in similar fraction in P. Aeruginos and hence Fig. 3 is showing that the effects of the G-stack probes on the correlation surface of the chip is very high. The highest correlation value of P. Aeruginos is about 0.8 (Table 3 is showing the highest correlation value is 1 for group 10 and group 12 but they are not considered because there is only one probe in groups 10 and 12).

On the other hand, the number of G-stack probes is much smaller in E.Coli GeneChip. Although the E.Coli chip designs show a reduction in the number of G-stack probes, the Figure 4 is showing that the effects of the G-stack probes on the correlation surface of E.coli chip is also very high. The highest correlation value of E.coli chip is also about 0.8 (for all the G-stacks to be at position eleven).

From the contour plots in Fig. 3 (E.coli) and Figure 4 (Pseudomonas Aeruginosa), it is clear that the hybridization level of GeneChips of two bacterias are affected by the presence of guanine rich probe sequences.

## 4 CONCLUSION

Previously, it has been reported that Affymetrix GeneChips of different mammals are affected by the formation of G-quadruplex structures due to the presence of guanine rich nucleotide sequences. Both the raw data as well as summerised data of mammalian GeneChip are affected.

This paper presented the examination of GeneChip data of two bacterias for the effects of G-stack probes that include Escherichia Coli and Pseudomonas Aeruginosa. The results of the two bacteria GeneChips

are as consistent as the results of mammalian GeneChips. Both the bacteria chips are showing high correlation among probes having guanine runs.

It has been observed that the 24% probe sets in Pseudomonas Aeruginosa GeneChip are affected because they have atleast one G-stack probe. Whereas, Escherichia Coli GeneChip has only 2% probe sets with G-stack probes in them. It is also noted that the chip size of Pseudomonas Aeruginosa is smaller than that of E.coli but the former has a higher fraction of G-stack probes. It is therefore assumed that either the genome of E.coli might not have frequent occurrences of guanine runs or it is by chance that GeneChip of E.coli has small fraction of G-stack probes. However, despite of the different fraction of G-stack probes in two bacteria GeneChips, the effects of G-stack can be seen for the both GeneChips.

**REFERENCES:**

Bates P., J. Mergny and D. Yang (2007) Quartets in G-major, EMBO reports, 8, 1003-1010.

Burge S., G. Parkinson, P. Hazel, A. Todd and S. Neidle (2006) Quadruplex DNA: sequence, topology and structure, Nucleic Acids Research, 34(19), 5402-5415.

Fasold M. and H. Binder (2012) Affy RNAD egradation: Control and correction of RNA quality effects in GeneChip expression data, Bioinformatics,29(1), 129-131.

Fasold M., H. Binder (2014) Variation of RNA quality and quantity are major sources of batch effects in microarray expression data, Microarrays, **3**(4), 322-339.

Huppert J. and S. Balasubramanian (2005) Prevalence of quadruplexes in the human genome, Nucleic Acids Research, 33(9), 2908-2916.

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(3), 259-277.

Marky L.A., H.T. Lee and A. Garcia (2010) Watson-Crick Base Pairs and Nucleic Acids Stability, eLS, Wiley Online Library.

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), 12-20.

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, Hindawi Journal of Nucleic Acids special issue on G-Quadruplex Nucleic Acids,

Memon F. N., Z. U. A. Khuhro, A. P. Harrison (2015) Affymetrix GeneChip of Arabidopsis Thaliana shows less effects of G-stack probes, Sindh University Research Journal (Science Series) ISSN 1813-1743, 47(3), 623-630.

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(1), 011906
(DOI:https://doi.org/10.1103/PhysRevE.68.011906).

Qin Y., L.H. Hurley (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplex found in eukaryotic promoter regions, Biochimie, 90(8), 1149-1171.

Siddiqui-Jain A., C.L. Grand, D.J. Bearss and L.H. Hurley (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, Proceedings of the National Academy of Science of the United States of America, 99(18), 11593-11598.

Shanaham H., F. N. Memon, G. Upton, A. Harrison (2011) 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(1), 13.

Upton Graham 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, Briefings in Functional Genomics and Proteomics, 8(3), 199-212.

Upton G., W. Langdon, A. Harrison (2008) G-spots cause incorrect expression measurement in Affymetrix microarrays, BMC Genomics, 9(1), 613.

Watson J.D. and F.H.C. Crick (1953) Molecular Structure of Nucleic Acids, Nature, 171(4356), 737-738.

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.