SAGE-Hindawi Access to Research
Journal of Nucleic Acids
Volume 2010, Article ID 489736, 6 pages
doi:10.4061/2010/489736

Research Article

A Comparative Study of the Impact of G-Stack Probes on Various Affymetrix GeneChips of Mammalia

Farhat Naureen Memon, Graham J. G. Upton, and Andrew P. Harrison

Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK

Correspondence should be addressed to Farhat Naureen Memon, fnmemo@essex.ac.uk

Received 1 December 2009; Revised 11 February 2010; Accepted 22 April 2010

Academic Editor: Jean Louis Mergny

We have previously discovered that probes containing runs of four or more contiguous guanines are not reliable for measuring gene expression in the Human HG_U133A Affymetrix GeneChip data. These probes are not correlated with other members of their probe set, but they are correlated with each other. We now extend our analysis to different 3′ GeneChip designs of mouse, rat, and human. We find that, in all these chip designs, the G-stack probes (probes with a run of exactly four consecutive guanines) are correlated highly with each other, indicating that such probes are not reliable measures of gene expression in mammalian studies. Furthermore, there is no specific position of G-stack where the correlation is highest in all the chips. We also find that the latest designs of rat and mouse chips have significantly fewer G-stack probes compared to their predecessors, whereas there has not been a similar reduction in G-stack density across the changes in human chips. Moreover, we find significant changes in RMA values (after removing G-stack probes) as the number of G-stack probes increases.

1. Introduction

A sequence of nucleotides having frequent occurrences of runs of guanines is capable of forming unusual four-stranded structures called G-quadruplex structures. These are a result of the Hoogesteen hydrogen bond that binds two guanines at close to 90 degrees. These structures not only form in a single sequence of nucleotides but two or four parallel sequences can also collectively form a G-quadruplex structure. We are investigating the effect of these unusual structures on gene expression measurements using microarrays.

Microarray technology is an effective way of gene expression profiling. Affymetrix GeneChips are the most popular type of array for many model organisms. GeneChip arrays are composed of 25-base long sequences that are known as probes. These probes are arranged in the form of pairs; each pair consists of a perfect match (PM) probe and a mismatch (MM) probe. The PM probe contains the same sequence of bases as appears in the gene; whilst MM is identical to PM except that the central base (13th position base) is the complement of that in the PM probe. Each probe belongs to a particular probe set and similarly each probe set represents a particular gene; however, some genes are represented by more than one probe set.

As a probe set corresponds to a particular gene, it is therefore expected that all the probes of a probe set should be correlated with each other if the particular gene that is represented by that probe set is expressed. However, [1] previously discovered that probes containing G-stacks behaved abnormally with respect to other probes in their probe sets. In our previous work [2], we confirmed the findings of [1], but we also went further in discovering that the G-stack probes are highly correlated with each other. This indicates that they cannot be measuring gene expression but instead suggests a biophysical process occurring on the surface of GeneChips, which we associate with the formation of G-quadruplexes. The probes on a GeneChip are grown through photolithography and this results in many single-stranded DNA sequences being held in close proximity [3]. Probes are readily able to physically touch their neighbouring probes, each of which shares the same sequence. It is expected that if these closely placed parallel probes contain runs of guanines then they may form G-quadruplex structures. In [2], we have also shown that

the value of the correlation coefficient changes according to the location of G-stack within the probes (using a popular human chip, the HG_U133A array).

We are now investigating different GeneChip designs for two issues.

(1) The effect of the position of guanine run within the G-stack probes with an expectation that G-stack probes are highly correlated with each other. We provide a detailed discussion on this topic in Section 3.1.

(2) The position of the G-stack with the highest correlation coefficient value with an expected result that this will be position 1. Position 1 is at the free end of the probe so it can more readily come into contact with its neighbouring probes (discussed in Section 3.2).

For this study we have selected chip designs that are used to study the transcriptome of the mammalian family. We have generated contour plots to show overviews of the entire correlation surface for each of these chip designs.

2. Materials and Methods

The GeneChip data consist of CEL files that report the average intensity of each probe of a microarray. These fluorescent intensities are read through the Affymetrix scanners after the target sequences are hybridised to a microarray. The data from many tens of thousands of GeneChip arrays are freely available in public domains in the form of CEL files. We have downloaded CEL files from the NCBI GEO (Gene Expression Omnibus) repository [4].

We have focused on the mammalian family and have selected GeneChip data for Homo Sapiens (Human), Mus Musculus (Mouse), and Rattus Norvegicus (Rat). For each organism, three or more different chip designs have been used. We have used data from 352 randomly chosen CEL files for most of the chip designs except for a mouse chip design MG_U74Bv2 (280 CEL files) and two of the human chips, HG_U95D (87 CEL files), and HG_U95E (86 CEL files).

We have adopted a pipeline for which a number of in-house informatics tools have been developed. The pipeline performs the following tasks.

2.1. To Generate Contour Plots

(1) We selected probes having exactly one G-stack of length four from the probe sequence (.tab) file of the particular chip design. The.tab file, which contains the probe annotation which includes probe set ID, x and y coordinates and probe sequence with some other information is available at the Affymetrix website [5].

(2) We separated out the filtered list of G-stack probes into groups according to the position of the guanine-run (G-stack) within the G-stack probes. The possible position of a G-stack having exactly four guanines within a probe could be $P = 1, 2, 3, \ldots, 22$. In this way, we have generated 22 groups of G-stack probes. For instance, group 1 represents to all the G-stack probes in which G-stack is at position one.

(3) Rather than using the observed intensities, we used the normalised CEL files.

(4) We produced lookup tables of the x and y coordinates for each of the 22 groups.

(5) As we have generated 22 groups of G-stack probes, a 22 by 22 matrix (M) is generated in which each element represents the average correlation coefficient of two groups of probes; probes that are members of one specified group with probes that are members of another specified group. For instance, element $M[5, 12]$ of the matrix represents the average correlation between G-stack probes in groups 5 and 12.

(6) As a final step, the matrix M is used to generate a contour plot of the correlation surface.

2.2. To Analyse RMA Values with the G-Stack Probes Included and Excluded

(1) In R with Bioconductor, we used RMA [6] to obtain a set of values for each probe set in each of 352 CEL file using the standard CDF file of the specific chip. This set of RMA values reflects to the values of the probe sets with the G-stack probes included.

(2) We then masked all the G-stack probes using the code supplied by NASC [7] in order to generate a new CDF file without G-stack probes.

(3) We again used RMA with the new CDF file to obtain another set of values of the probe sets with the G-stack probes masked.

(4) These two sets of RMA values were used to analyse the effect of removing G-stack probes on RMA values (discussed in Section 3.3).

3. Result and Discussion

A list of chip designs involved in this study is shown in Table 1. The table also shows the chip size, the number of annotated probes, the number of G-stack probes, and the number of affected probe sets in each chip design. The lists of G-stack probes and the lists of affected probe sets (along with the number of G-stack probes in those probe sets) for the arrays analysed are available at http://bioinformatics.essex.ac.uk/users/fnmemo/G-Tract.html.

3.1. Effect of the Position of G-Stack within the Probes. The contour plot for human chip HG_U133A, Figure 1, is almost identical to our previous work, with the discrepancy arising because of the different datasets used in the two studies. The density of G-stack probes differs according to the position of G-stack within the probes (see Table 2). The correlation coefficients of G-stack probes in all the human chips are quite high with the most marked correlation values at their diagonals.

Human arrays show a fairly constant fraction of G-stack probes across different designs. For instance, for

Table 1: List of organisms and their chip designs used in this study. The number of annotated probes and the number of G-stack probes include both the main and control probes.

Organism	Chip Design	Chip size	No. of annotated Probes	No. of G-stack Probes	No. of Affected Probe Sets
Humans	HG_U133_Plus_2	1164 ∗ 1164	604,258	24,980	16,254
(Homo_Sapiens)	HG_U133A	712 ∗ 712	247,965	12,868	8,298
	HG_U95A	640 ∗ 640	201,807	7,329	3,733
	HG_U95B	640 ∗ 640	201,862	6,334	3,240
	HG_U95D	640 ∗ 640	201,858	7,198	3,227
	HG_U95E	640 ∗ 640	201,863	7,880	3,514
Mouse	MOE430A	712 ∗ 712	249,958	372	314
(Mus Musculus)	MOE430B	712 ∗ 712	248,704	252	203
	MG_U74Av2	640 ∗ 640	197,993	7,360	3,556
	MG_U74Bv2	640 ∗ 640	197,131	7,006	3,614
Rat	RAE230A	602 ∗ 602	175,477	81	58
(Rattus Norvegicus)	Rat230_2	834 ∗ 834	342,410	208	163
	RG_U34A	534 ∗ 534	140,317	3,691	2,104

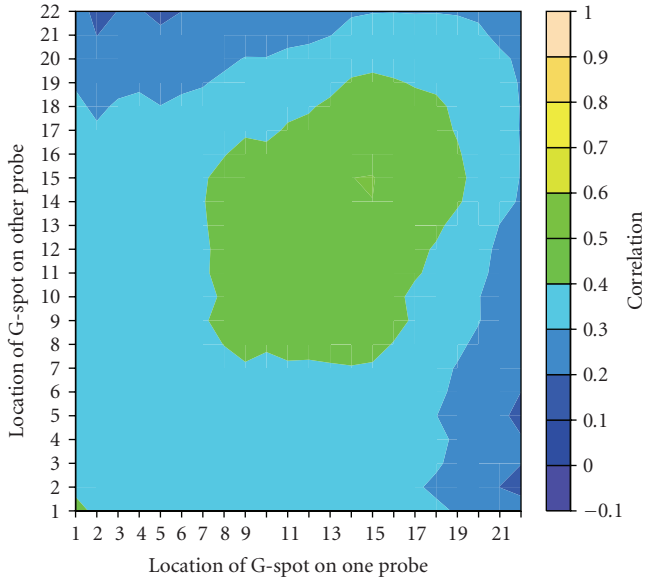

Figure 1: Contour plot illustrating that in human chip HG_U133A, the average correlation coefficient values changes according to the position of G-stack (with four Gs only) for a group of probes.

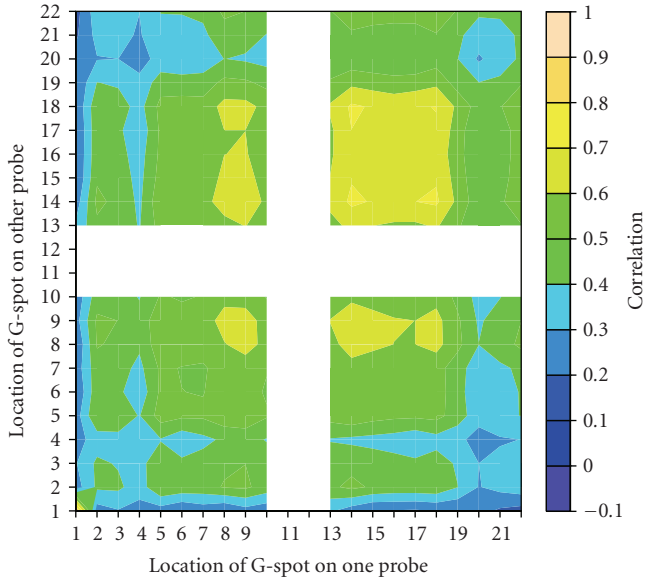

Figure 2: Contour plot illustrating that in mouse chip MOE430B, the average correlation coefficient values changes according to the position of G-stack (with four Gs only) for a group of probes.

a recent human chip design, HG_U133_Plus_2, 4.1% (24,980/604,258) of the annotated probes contain a G-stack. On the older HG_U95A design 3.6% (7,329/201,807) of the annotated probes contain a G-stack.

In contrast to the human results, the mouse and rat chips show large changes in G-stack probe density across different designs, with significantly smaller fractions being found on the latest designs. Of the four mouse chips used in this study, the older designs MG_U74Av2 and MG_U74Bv2 have over an order of magnitude more G-stack probes than do the newer chip designs, MOE430A and MOE430B. We found that 0.15% (372/249,958) and 0.1% (252/248,704) of the annotated probes contain a G-stack in MOE430A and MOE430B, respectively. Whereas the percentages of annotated G-stack probes in MG_U74Av2 and MG_U74Bv2 are 3.72% (7,360/197,993) and 3.6% (7,006/197,131), respectively. Moreover, both the chips, MOE430A and MOE430B, also have an absence of G-stack probes in their middle (see Table 2) which is reflected in the contour plot of chip MOE430B in Figure 2. As with the human data, the locations for which there are probes indicate that G-stacks probes are highly correlated on mouse designs. Furthermore, the correlation is also highest when comparing two probes that have G-stacks at the same location within the probe.

We also used three chip designs for rat (Rattus Norvegicus). The RG_U34A chip is the oldest design and we find

Table 2: In G-stack probes, the effect of the position of G-stack on the average correlation coefficient value, (n is the number of affected probes and $\bar{r}$ is the average correlation between these n probes).

		Position of G-stack										
		1	2	3	4	5	6	7	8	9	10	11
HG_U133A	n	871	449	548	583	664	546	599	604	531	471	458
	$\bar{r}$	0.51	0.32	0.36	0.36	0.34	0.34	0.36	0.40	0.44	0.42	0.44
HG_U133_Plus_2	n	1758	925	1072	1170	1234	1087	1214	1093	1049	923	903
	$\bar{r}$	0.29	0.15	0.17	0.18	0.19	0.18	0.18	0.23	0.26	0.22	0.27
HG_U95A	n	398	297	255	271	315	308	367	448	417	47	47
	$\bar{r}$	0.40	0.28	0.30	0.31	0.31	0.29	0.33	0.35	0.40	0.42	0.39
HG_U95B	n	314	267	237	261	282	293	326	375	339	19	23
	$\bar{r}$	0.66	0.51	0.54	0.55	0.56	0.57	0.58	0.64	0.67	0.63	0.68
HG_U95D	n	293	278	243	272	284	311	381	471	478	56	73
	$\bar{r}$	0.60	0.36	0.37	0.39	0.38	0.41	0.42	0.44	0.50	0.41	0.47
HG_U95E	n	346	323	317	285	350	363	398	532	559	73	57
	$\bar{r}$	0.54	0.29	0.32	0.35	0.34	0.33	0.41	0.41	0.46	0.39	0.47
MOE430A	n	15	19	16	14	15	22	36	12	12	13	0
	$\bar{r}$	0.51	0.26	0.32	0.30	0.15	0.28	0.27	0.33	0.33	0.40	—
MOE430B	n	4	13	14	13	16	13	22	4	9	8	0
	$\bar{r}$	0.92	0.42	0.39	0.31	0.54	0.50	0.49	0.60	0.62	0.49	—
MG_U74Av2	n	357	315	259	292	292	349	349	428	431	39	46
	$\bar{r}$	0.29	0.15	0.17	0.18	0.19	0.18	0.18	0.23	0.26	0.22	0.27
MG_U74Bv2	n	326	286	246	257	271	298	369	436	496	17	14
	$\bar{r}$	0.54	0.33	0.33	0.35	0.39	0.40	0.42	0.46	0.54	0.61	0.63
RG_U34A	n	194	148	145	126	166	160	183	176	144	94	85
	$\bar{r}$	0.33	0.23	0.21	0.24	0.28	0.29	0.32	0.32	0.41	0.38	0.36
		Position of G-stack										
		12	13	14	15	16	17	18	19	20	21	22
HG_U133A	n	491	424	523	580	592	604	650	615	737	611	689
	$\bar{r}$	0.45	0.47	0.50	0.50	0.47	0.43	0.41	0.38	0.34	0.29	0.26
HG_U133_Plus_2	n	949	872	1009	1140	1098	1193	1271	1185	1310	1192	1308
	$\bar{r}$	0.29	0.32	0.39	0.42	0.39	0.41	0.39	0.38	0.32	0.27	0.24
HG_U95A	n	33	37	631	417	384	388	359	433	467	399	570
	$\bar{r}$	0.42	0.49	0.54	0.55	0.55	0.57	0.55	0.57	0.55	0.54	0.54
HG_U95B	n	15	18	555	336	350	342	340	383	390	365	463
	$\bar{r}$	0.77	0.84	0.81	0.79	0.81	0.81	0.81	0.80	0.77	0.78	0.75
HG_U95D	n	72	74	832	364	357	337	330	391	426	334	500
	$\bar{r}$	0.43	0.54	0.64	0.64	0.66	0.75	0.69	0.70	0.74	0.71	0.75
HG_U95E	n	61	73	833	412	378	371	354	420	434	370	530
	$\bar{r}$	0.39	0.47	0.62	0.61	0.66	0.64	0.67	0.68	0.67	0.69	0.65
MOE430A	n	2	25	16	15	27	20	19	18	8	14	7
	$\bar{r}$	0.37	0.33	0.40	0.42	0.37	0.36	0.26	0.29	0.31	0.10	0.07
MOE430B	n	0	7	10	19	23	12	10	10	6	6	6
	$\bar{r}$	—	0.55	0.72	0.66	0.62	0.68	0.70	0.45	0.28	0.32	0.48
MG_U74Av2	n	42	51	682	434	388	373	362	457	471	357	545
	$\bar{r}$	0.29	0.32	0.39	0.42	0.39	0.41	0.39	0.38	0.32	0.27	0.24
MG_U74Bv2	n	6	11	790	404	392	336	326	443	419	357	465
	$\bar{r}$	0.78	0.54	0.68	0.67	0.67	0.69	0.62	0.62	0.57	0.56	0.55
RG_U34A	n	82	76	182	185	214	200	204	215	213	202	268
	$\bar{r}$	0.47	0.43	0.46	0.49	0.46	0.42	0.42	0.45	0.34	0.26	0.21

Table 3: The effect on RMA of removing G-stack probes from probe sets. Subtraction of the original RMA value from the RMA value after removal of G-stack probes gives the quantity d. Entries are column percentages.

No. of G-stack probes:	0	1	2	3	4	5	6
No. of probe sets:	38,422	10,216	4,192	1,280	384	121	36
$d > 2.0$	0	0	0	0	0	0	2
$1.0 < d \leq 2.0$	0	0	1	1	3	3	9
$0.5 < d \leq 1.0$	0	1	4	5	7	6	11
Between −0.5 and 0.5	100	94	76	61	51	46	34
$-0.5 > d \geq -1.0$	0	5	14	21	19	19	10
$-1.0 > d \geq -2.0$	0	0	5	11	17	21	18
$d < -2.0$	0	0	0	1	2	5	16

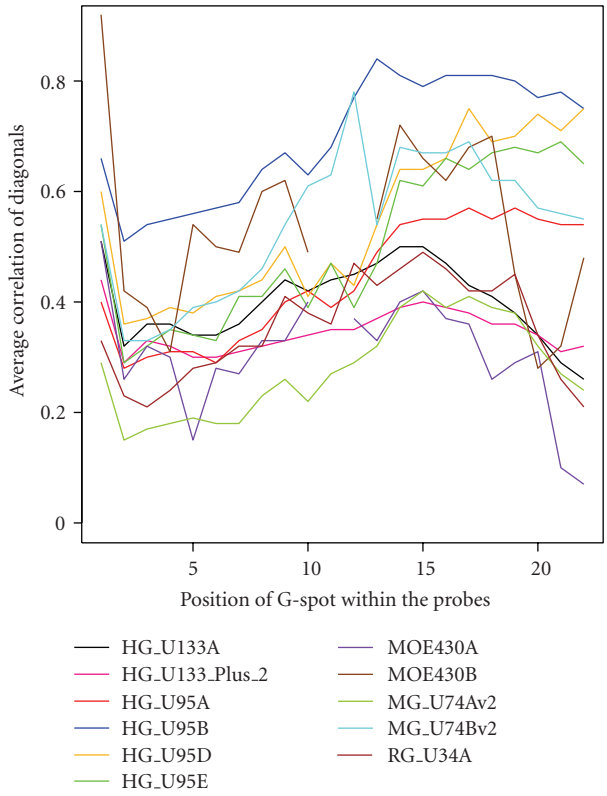

Figure 3: The plot shows that diagonal correlation coefficient values of each chip design. Diagonal values represent correlation among the same group of probes.

that it has over an order of magnitude more G-stack probes than do the new chip designs, Rat230_2 and RAE230A. We found 0.05% (81/175,477) of the annotated probes contain a G-stack in chip design RAE230A. Similarly chip design Rat230_2 has 0.06% (208/342,410) of the annotated probes contain a G-stack. Whereas, the old Rat chip RG_U34A contains 2.6% G-stack probes (3,691/140,317).

Due to the small number of G-stack probes in the new chips of rat and mouse, we expect that the gene expression measurement in these new chip designs will be less affected by G-stack probes.

3.2. Position of the G-Stack with the Highest Correlation Coefficient Value. We have also checked whether there is any specific position of G-stack where the correlation is always most marked. We were expecting that the correlation could be highest at the 5′ end of the probes, as is the case for HG-U133A in our previous work [2]. The 5′ end of the probe is free and so it has a greater tendency to come into physical contact with adjacent probes. However, we find that there is no specific position of G-stack where correlation coefficient is most marked in all the chips. Furthermore, the diagonal values in the contour plot are almost always showing the highest correlation. Table 2 provides the details of diagonal values of matrix M (M is explained in Section 2.1) for all the chips analysed which is graphically illustrated in Figure 3.

3.3. Effect of Removing G-Stack Probes on RMA Values. To examine the effect of removing G-stack probes, we used RMA to obtain values for each probe set in each of 352 HG_U133_Plus_ 2 CEL files. We then obtained revised RMA values with the G-stack probes masked. In Table 3, we report the results in terms of a summary of the values obtained for d, which we define as the revised RMA value minus the original value. In the table there are separate columns to summarise the effects on probe sets having varying numbers of G-stack probes. The percentages reported are based on the number of probe sets shown in the second row of the table and are averaged over the 352 CEL files.

As one would hope, there are no major changes on probe sets that have no G-stack probes. As the number of G-stack probes increases, so the changes become potentially much more serious, and the effects are more variable. On average G-stack probes have higher values than other probes, so that, the majority of RMA values are reduced by the removal of the G-stack probes. However, there are also many instances where the RMA value is appreciably increased by removal of G-stack probes.

4. Conclusion

G-stack probes behave as outliers within their probe sets because they are usually poorly correlated with other members of their probe sets while they are highly correlated with each other. We have illustrated that this is true in

various chip designs of different mammalia. Therefore, as we suggested before in our previous work [2], these probes should not be included within a calculation of the gene expression measurement. Due to the previous work, we were expecting that the correlation among the G-stack probes is at its highest when the runs of guanines start from position 1 (5′ end) within the probes. It was our expectation that as the 5′ end is the free/moving end, so there are more changes for the G-stack probes to attach with the neighbouring probe's G-stack at this end. Although it is true for some chip designs, for instance HG_U133A, MOE430A and MOE430B, it is not true for all of them. Thus, in general, we did not find a common position of G-stack where the correlation coefficient value is high in all the chips. We also found that a much smaller fraction of G-stack probes are present in the new chip designs of rat and mouse compared to the original designs. This suggests that the change in design protocol led to a significant removal of probes which we now believe to be misinformative. It is surprising that such a change in design did not lead to a significant reduction in the amount of G-stack probes in human 3′ array.

Furthermore, we find that the changes in RMA values (after removing the G-stack probes) become more serious as the number of G-stack probes increases.

Acknowledgments

Memon is funded by a scholarship from the University of Sindh (NO.SU/PLAN/F.SCH/611). The authors are grateful to Dr. Neil S. Graham and Dr. Olivia Sanchez-Graillet for help with software.

References

[1] C. Wu, H. Zhao, K. Baggerly, R. Carta, and L. Zhang, "Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays," *Bioinformatics*, vol. 23, no. 19, pp. 2566–2572, 2007.

[2] G. J. G. Upton, W. B. Langdon, and A. P. Harrison, "G-spots cause incorrect expression measurement in Affymetrix microarrays," *BMC Genomics*, vol. 9, article 613, 2008.

[3] W. B. Langdon, G. J. G. Upton, and A. P. Harrison, "Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 259–277, 2009.

[4] T. Barrett, T. O. Suzek, D. B. Troup et al., "NCBI GEO: mining millions of expression profiles - Database and tools," *Nucleic Acids Research*, vol. 33, pp. D562–D566, 2005.

[5] Affymetrix, http://www.affymetrix.com/index.affx.

[6] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[7] NASC's International Affymetrix Service, http://affymetrix.arabidopsis.info/xspecies/.