Detectability on RNA expression microarrays

Sigrun Helga Lund,
Gunnar Stefansson.

RH-16-2010

Science Institute
University of Iceland
Dunhaga 3, 107 Reykjavik

August 2010

Contents

Abstract

Tiled microarrays are widely used to assess expression along the genome. Usually, the same probes are used for all microarrays which contain samples from different individuals. Therefore will SNP's cause some probes not to match perfectly. By using repeated copies of probe pairs, which only differ by a single base, we can estimate the effect on RNA expression of altering a single base at the middle of a probe. Power analysis showed that given enough repetitions around half of probe-pairs have significantly different intensity levels when a single base is altered. It is also shown that perfectly matching probes can have higher mean intensity levels than mismatching probes.

1 Introduction

Tiled microarrays are widely used to assess expression of protein-coding genes, non-coding RNAs and transcription in "gene deserts" [1, 2]. The arrays contain probes representing a target genomic region, which is referred to as tiling the region. Signal intensities of the probes along the region are then used to estimate the amount of RNA transcription at the corresponding cites.

Application of microarrays is technically challenging due to several sources of variation which significantly degrade the ability to accurately detect the true gene-expression signal [3].Several statistical methods are currently in use for taking technical variation into account [4, 5], some of which require a mismatch probe paired with each probe, which differs in sequence by only one base. When using tiling microarrays further issues arise, as different sets of biases arise as opposed to the differential expression analysis

Usually, tiling microarray experiments are done on several samples from several individuals with different genotypes. The arrays, on the other hand, are all identical, containing the same probes. As the probesets represent whole genomic regions, they unavoidably contain probes from genomic locations where SNPs are situated. As the samples used in the experiment have different genotypes, they will have different versions of these SNPs. One hypothesis is that this will cause bias in signal intensities as some probes will match perfectly to the genotype of some samples and others not.

For probes representing locations where the individual is homozygote in the corresponding SNP, the probe pair can be viewed as a pair of a perfect match and a mismatch probe. It has been noted that mismatch probes give higher signal intensities than perfectly matching probes in 30 % , but as microarray data is noisy, it is not clear whether that effect is caused by true differences in mean intensity levels, or solely the noise itself.

The main objective of this study is to investigate the amount of bias caused by mismatches due to SNPs. To that mean, the array contained 16

probe pairs, corresponding to the two alleles of 16 known SNP's. In order to achieve a good estimate of the true intensity levels, the array included repeated copies of each probe. The allele-to-allele variation is compared to, among other others, the genomic location-to-location variation and a power analysis is conducted to investigate how many repetitions of each probe pair are needed in order to call the allele-to-allele variation significantly different from zero. Finally, it is checked whether the difference in signal intensities from allele-to-allele is in concordance with the genotype of the individuals and meanwhile investigated whether perfectly matching probes give truly higher mean intensity levels than those including a single mismatch.

2 Methods

The data used in this paper is RNA expression data from 7 Nimblegen microarrays. The RNA was extracted from normal tissue of 7 different prostate cancer patients, one sample for each array. We confine our analysis to a part of the probeset for the experiment, consisting of repeated copies of sixteen pairs of 60 mer probes. The two probes in each probe-pair cover exactly the same location on the genome where known single nucleotide polymorphisms (SNP's) are located, all of which have been associated with prostate cancer. The two probes contain the two known alleles, which are situated at the middle of the probe, but all other bases are identical. We therefore have 32 different probes, each of which is repeated 96 times on the array.

In order to minimize spatial artifacts in the expression signal, the wells of the microarray were split up into 24 non-overlapping "containers", which formed a 4x6 grid on the array. Four copies of each pair of probes were allocated to each container. The location of probes within each container was subsequently randomized.

As (find ref) have pointed out normalization on microarrays should be carefully conducted as it can cause spurious correlations within the data. Therefore the only normalization applied was to first take the logarithm of the data and then subtract the median response of the probes within each container on each array.

3 Results

3.1 Mixed effects model

We fit the following mixed effects model

$$y_{acpsr} = \alpha_a + \beta_{c(a)} + \pi_p + (\alpha\pi)_{ap} + \gamma_{s(ap)} + \varepsilon_{acpsr},$$

where α_a are fixed array effects, $\beta_{c(a)}$ are random container effects, nested within the arrays, π_p are random location effects, representative for

variance comp.	estimate	perc. of total variance
σ^2_β	0.003	2.5 %
σ^2_π	0.059	48.8 %
$\sigma^2_{(\alpha\pi)}$	0.025	20.7 %
$\sigma^2_{\gamma(\alpha\pi)}$	0.011	9.1 %
σ^2	0.023	19.0 %

Table 1: Estimate of the variance components for the mixed effects model

each probe-pair, $(\alpha\pi)_{ap}$ are random array-location interaction effects, $\gamma_{s(ap)}$ are random base effects, nested within the array-location interaction and ε_{acpsr} are iid and normally distributed. Table 3.1 shows the estimates of the variance of the random effects. Not surprisingly, the location-to-location variation is by far the largest component, contributing almost half of the total variation. There is also a considerate array-location variation, 20.7 % of the total variation, showing that difference in signals by location varies between subjects. Notice that the base-to-base variation is almost half of the array-probe variation or 9.1 % of the total variation in expression levels. Although the error component is more than twice as large as the base component, the former one can be reduced by having repeated copies of each probe. If we have each probe repeated R times within each of the C containers, the variance of the difference in the expression levels of the probe-pair containing the two different bases is

$$\mathrm{Var}[\bar{y}_{a.ps_1.} - \bar{y}_{a.ps_2.}] = 2\left(\frac{\sigma^2_\beta}{CR} + \sigma^2_\gamma + \frac{\sigma^2}{CR}\right).$$

This shows that as soon as we have more than two repeated copies of each probe-pair, the base-to-base variation dominates the error.

Now the question arises how many repetitions of each probe are needed in order for the σ^2_γ component to be called significantly different from zero. That can be tested via a simple F-test, but whereas the residuals are far from being normal, the α-level of the test is not certain. Therefore we first conducted the F-test on 100000 bootstrap samples generated from the empirical null distribution (by randomly "assigning" bases to probe-pairs). The 0.05 quantile of the F-values gotten from that bootstrap are then used as the cutoff value for the latter bootstrapping on the original data. In both of the bootstrapping, each "pair" of probes on each array was repeated for the given number of repetitions. In order to test for as few as two repetitions a simpler model was assumed where there are no containers (nor container effects) assumed. The results are shown in Table 3.1. There it can be seen that when we have three repetitions, there is almost 50 % power for calling the base effect significant, and as soon as we have four repetitions of each probe, the power is at 97%.

repetitions	2	3	4	5
power	0.166	0.490	0.970	0.979

Table 2: Number of repetitions needed in order for σ_γ^2 to be significantly greater than zero. Here we conduct F-tests on bootstrapped samples, where we compare the bootstrapped F- values to a cutoff F-value, retrieved from bootstrapping of the empirical null distribution. The α-level is 0.05

3.2 What probes contribute the most variation?

It is interesting to see whether the base-to-base difference in expression levels is due to a single extreme probe-pair on a single array or a more global effect. Figure 3.2 gives a good overview for estimating due to which probe the difference is detected. For convenience we have labeled each probe-pair with a letter and each array with a color. We then plot the mean response levels of the probes containing one version of the variating base against the mean response levels of the probes containing the other version. Then a letter in the color corresponding to the given array is printed at the obtained coordinate. The figure reveals that the base-to-base difference is not solely due a single pair on a single array as pairs A, F, G, J and O seem to show differences in mean expression on several arrays.

Paralell boxplots of the signal intensities for both variations of the five pairs are plotted for each array in Figure 3.2. The difference is evident for probe-pair "F", on array 2, for example, the intensities are perfectly separated and there is also a undeniable difference on arrays 1,2,5 and 6. Probe pair "O" also shows considerable general difference, whereas the only obvious difference for probe-pair "G" is on array 2. Another thing to notice is the difference in variability from array to array. On array 5 there is a clear difference in signal intensities for all pairs, besides perhaps probe-pair "J", whereas the plot for array 3 doesn't reveal clear differences for any of the pairs.

That difference was confirmed by conducting the Welch's t-test for each prope-pair. As the difference varies greatly from location-to-location and also between arrays, the analysis was done for each of the five probe-pairs and each array separately. As before 100000 bootstrap samples of the empirical null (where bases are randomly 'allocated' to probes) are used to find a cutoff t-value, as the normal premise does not hold for the data.

Let's now take a closer look at the 5 locations where the corresponding probe-pair shows the most difference in signal intensities, that is locations A,F,G,J and O. A power analysis of how many repetitions are needed in order to call a difference in the expression levels of the two variants was conducted. A bootstrapped Welch's t-test was conducted as before, but now for 2,4, 10 and 96 repetitions. The results of the power analysis are

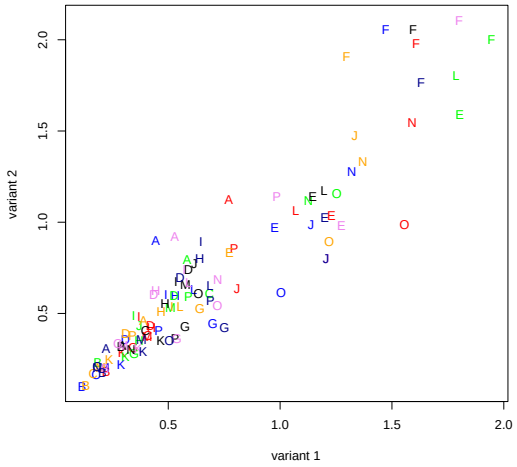

Figure 1: The mean expression levels of the probes containing one version of the variating base plotted against the mean response levels of the probes containing the other version. A letter corresponding to the probe-pair in a color corresponding to the array is printed at the obtained coordinate.

shown in Table 3.2. They are in good concordance with what was noticed in the parallel boxplots. With 10 repetitions of each probe, there is generally good power for detecting the difference in the probe-pair "F" and Array-3 has low power for all pairs. Notice that given a vast number of repetitions (96), 26 out of 35 probe-pairs have more than 80% power of calling a difference in mean intensity levels. When the repetitions are only 10, half of them have greater than 80 % power, and when the repetitions are only four, 6 probe-pairs have gained the desired power.

3.3 Agreement with genotype?

As we have seen, we can detect a difference in signal intensities of two probes, which are identical except for one base, situated in the middle of the probe. As the alternating bases all correspond to known SNP's, it is of interest to see whether the difference in expression is in concordance with the genotype of the individuals. In this study, the genotype was previously known for 89 of the $7 \cdot 16 = 112$ cases. In table 4 these 89 cases are categorized by a) whether the sample is homozygote or heterozygote for the corresponding SNP, b) whether the corresponding probe-pair has significantly different signal intensities. There is no evident relationship between the genotype (homeo/heterozygote) and whether the probe pair has significantly different signal intensities (p=0.65).

Let us now take a closer look at those 32 cases where the probe-pair

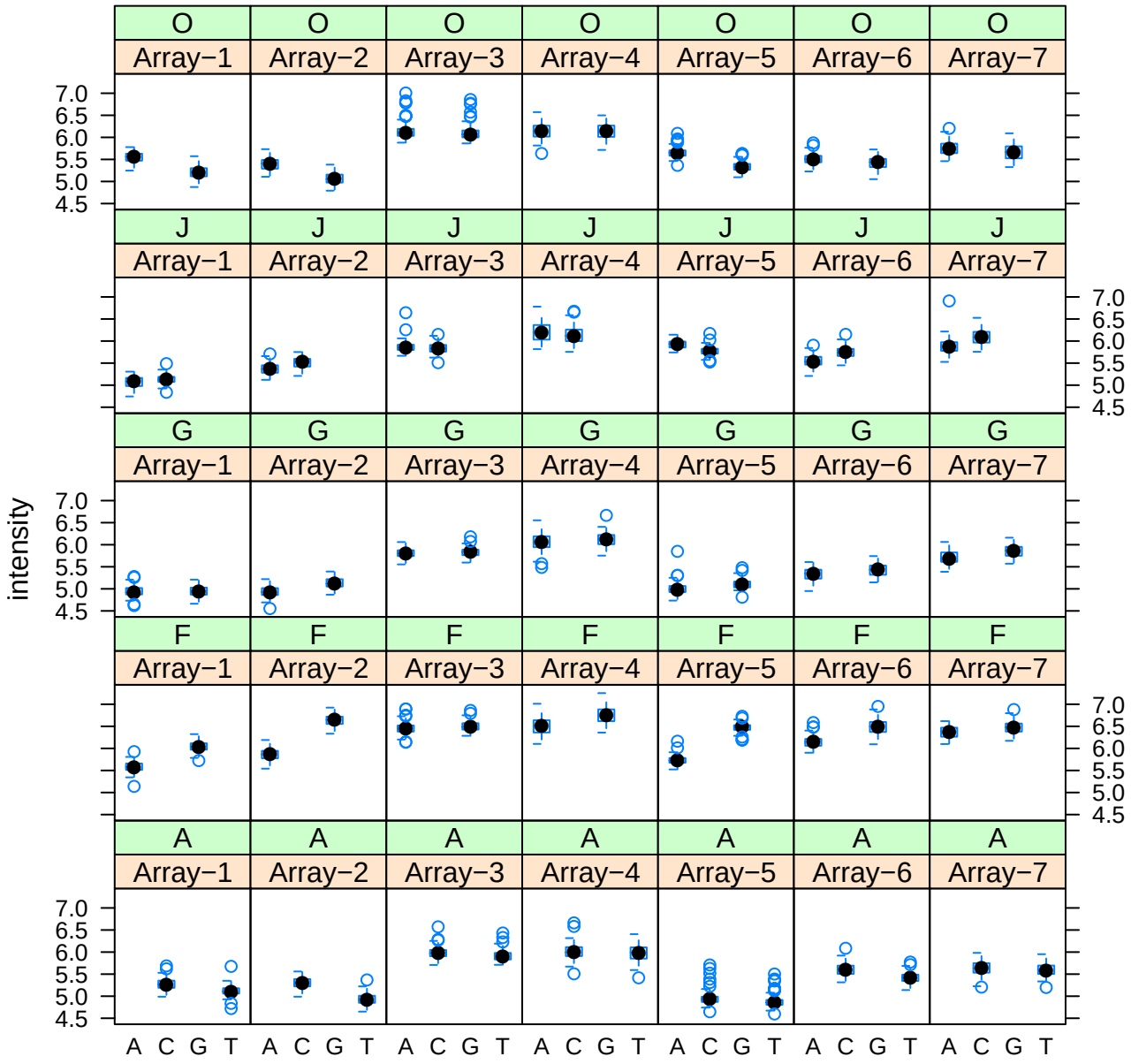

Figure 2: Parallel boxplots of the signal intensities for each probe on each array. The probe pair is denoted in brackets, followed by the variating base.

Pair	reps	Array-1	Array-2	Array-3	Array-4	Array-5	Array-6	Array-7
A	2	0.13	0.31	0.07	0.05	0.08	0.14	0.05
	4	0.37	0.92	0.12	0.05	0.13	0.42	0.06
	10	0.81	1.00	0.23	0.05	0.20	0.89	0.08
	96	1.00	1.00	0.95	0.21	0.85	1.00	0.45
F	2	0.36	0.40	0.05	0.12	0.38	0.25	0.07
	4	0.99	1.00	0.07	0.34	1.00	0.77	0.15
	10	1.00	1.00	0.09	0.83	1.00	1.00	0.42
	96	1.00	1.00	0.46	1.00	1.00	1.00	1.00
G	2	0.00	0.02	0.01	0.01	0.01	0.01	0.01
	4	0.00	0.05	0.01	0.01	0.03	0.01	0.02
	10	0.01	0.10	0.01	0.01	0.06	0.03	0.06
	96	0.11	1.00	0.39	0.89	1.00	1.00	1.00
J	2	0.06	0.10	0.05	0.05	0.16	0.14	0.12
	4	0.10	0.25	0.05	0.06	0.54	0.43	0.35
	10	0.22	0.64	0.07	0.10	0.93	0.90	0.77
	96	0.98	1.00	0.40	0.61	1.00	1.00	1.00
O	2	0.26	0.23	0.05	0.05	0.30	0.06	0.06
	4	0.83	0.79	0.06	0.05	0.89	0.11	0.09
	10	1.00	1.00	0.08	0.04	1.00	0.28	0.19
	96	1.00	1.00	0.29	0.06	1.00	1.00	0.95

Table 3: The power for calling a difference in the expression levels of the two variants of each probe-pair for each array specifically.

	significant difference	not significant
homozygote	32	27
heterozygote	14	16

Table 4: The cases when the genotype of the sample for the corresponding SNP is known listed by whether the probe pair had significantly different signal intensities and genotype.

has significantly different expression levels, and the corresponding sample is homozygote for that particular SNP, In 28 out of these 32 cases, the probe matching perfectly to the genotype had greater signal intensities than the other one, but in 4 cases, the probe with a single mismatch gave significantly greater intensity levels.

4 Discussion

In this study it is shown that in 57 out of 112 cases, probe pairs that only differ by a single base have significantly different signal intensities, given enough repetitions of the probes.

When there are more than two repetitions of each probe, the allele-to-allele variation dominates the error. Only four repetitions are needed to gain 97% power for calling the allele-to-allele variation significantly greater than zero in a random effect model.

It is not certain that the probe matching perfectly will give higher signal intensities than the one with a single mismatch. The reason for mismatch probes giving higher intensity levels is not solely caused by general variability of the data, mismatching probes can truly give higher expression levels.

References

[1] J.M. Johnson, S. Edwards, D. Shoemaker, and E.E. Schadt. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *TRENDS in Genetics*, 21(2):93–102, 2005.

[2] T.C. Mockler and J.R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85(1):1–15, 2005.

[3] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *nature genetics*, 32(supp):490–495, 2002.

[4] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

[5] T.E. Royce, J.S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics*, 21(8):466–475, 2005.