FEBS Letters 581 (2007) 3363–3370

Identification of an in vitro transcription-based artifact affecting oligonucleotide microarrays

David C. Nelson[a,1], Dana J. Wohlbach[a], Matthew J. Rodesch[a], Viktor Stolc[b], Michael R. Sussman[a,*], Manoj P. Samanta[c,1]

[a] *Biotechnology Center and Biochemistry Department, University of Wisconsin, 425 Henry Mall, Madison, WI 53706, United States*
[b] *NASA Ames Genome Research Center, Moffet Field, CA, United States*
[c] *Systemix Institute, Los Altos, CA, United States*

Received 3 March 2007; revised 24 May 2007; accepted 14 June 2007

Available online 26 June 2007

Edited by Frances Shannon

Abstract **This study identified the widely used T7 in vitro transcription system as a major source of artifact in the tiling array data from nine eukaryotic genomes. The most affected probes contained a sequence motif complementary to the +1 to +9 initial transcribed sequence (ITS) of the T7-$(dT)_{24}$ primer. The abundance of 5′ ITS cRNA fragments produced during target preparation was sufficient to drive undesirable hybridization. A new T7-$(dT)_{24}$ primer with a modified ITS was designed that shifts the artifactual motifs as predicted and reduces the effect of the artifact. A computational algorithm was generated to filter out the likely artifactual probes from existing whole-genome tiling array data and improve probe selection. Further studies of *Arabidopsis thaliana* were conducted using both T7-$(dT)_{24}$ primers. While the artifact affected transcript discovery with tiling arrays, it showed only a minor impact on measurements of gene expression using commercially available 'gene-only' expression arrays.**

Keywords: RNA; Microarrays; Genes; Chips; DNA; Expression

1. Introduction

Microarray technology has provided unprecedented opportunities to study an entire genome in a high throughput manner. As recent developments in chip production permit higher oligonucleotide density and more versatile chip designs, applications have extended beyond the analysis of annotated gene expression to include unbiased whole-genome exploration [1–11]. This has led to improved detection of predicted intron/exon boundaries and transcriptional splice variants, as well as the identification of many putative small non-coding RNAs and miRNAs. A number of unexpected transcriptional phenomena, such as widespread signals in antisense, intronic and intergenic regions, were also observed in whole-genome tiling array experiments performed in several organisms. Although the biological significance of this unusual transcription remains largely unknown, an even more critical issue is which of the observed signals are a genuine reflection of in vivo transcription. As microarrays can be very sensitive to low levels of transcription, it is not always possible to verify the authenticity of a putative transcript through traditional methods involving Northern blots or reverse transcriptase polymerase chain reaction (RT-PCR). The large number of detected RNAs would also make the verification process costly and laborious. As the number of probes used in microarray investigations continues to rise, even a low false positive rate of detection can be expected to produce increasingly significant errors in genome discovery. Consequently it is of utmost importance to identify sources of artifact and take steps to reduce the false positives that inevitably arise in this technology [12–14].

It has previously been noted that experimental artifacts may arise through cross-hybridization, contamination from genomic DNA and unspliced RNA or unintended double-stranded labeling of RNA [13]. In addition, it is expected that different microarray platforms and labeling systems will have a set of artifacts specific to the methods used (see Supplementary Discussion). This study provides evidence of an important sequence-specific artifact for all oligonucleotide microarrays that use T7 RNA polymerase for cRNA target production. Technical and computational solutions to ameliorate this problem are also presented.

2. Results

2.1. High signal bias in C-rich probes

A comparison of the signal intensity and nucleotide composition of probes was performed on data from previously published *Arabidopsis thaliana* 36-mer whole-genome tiling microarrays synthesized on a maskless photolithography platform [4,15]. In contrast with the entire set of oligos, the top 5% of probes ranked by signal intensity showed a marked increase in cytosine and decrease in adenine content, while the frequency of thymine and guanine residues remained relatively unaffected (Fig. 1). This trend was reproducible in data from an earlier *Arabidopsis* tiling study using 25-mer Affymetrix arrays, indicating a common bias that was independent of probe length, chip synthesis technology, or laboratory handling [2]. Further analyses were performed on data from other eukaryotic arrays, and a similar bias was noted for honeybee

*Corresponding author. Fax: +1 608 262 6748.
E-mail address: msussman@wisc.edu (M.R. Sussman).

[1]These authors contributed equally to this work.

0014-5793/$32.00
doi:10.1016/j.febslet.2007.06.033

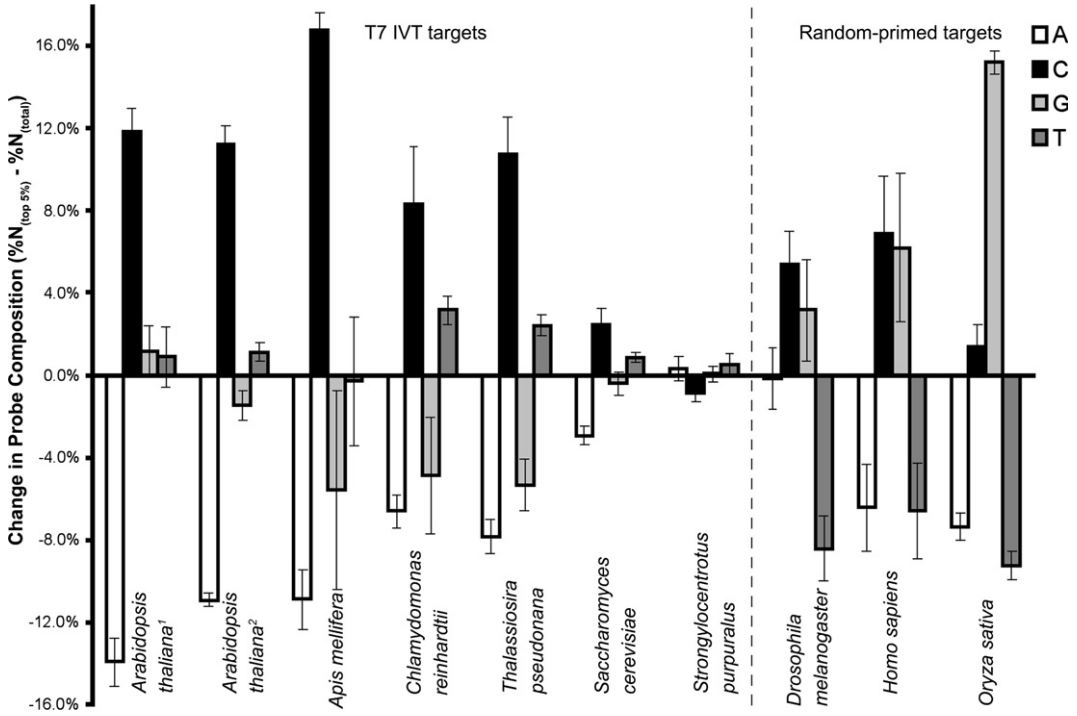

Fig. 1. Nucleotide composition shifts among high signaling probes. The nucleotide abundances among the top 5% highest signaling oligos from an individual array were compared with the overall nucleotide abundances among all probes on that array. The average change in probe composition for multiple arrays (48, 13, 17, 6, 17, 8, 27, 24, 135, and 90 arrays, respectively) from each genome is shown. Error bars indicate standard deviation. *A. thaliana*[2] [4], *A. mellifera*, *C. reinhardtii*, *T. pseudonana*, *S. cerevisiae*, *S. purpuratus* tiling arrays were hybridized with cRNA derived from T7-based IVT reactions. *D. melanogaster*, *H. sapiens*, and *O. sativa* tiling arrays were hybridized with randomly primed first strand cDNA. *A. thaliana*[1] [2] arrays were composed of 25-mer probes for the Affymetrix platform; *S. purpuratus* was a 50-mer array synthesized by maskless photolithography; all other arrays had 36-mer probes also synthesized with maskless photolithography.

(*Apis mellifera*), *Chlamydomonas reinhardtii*, and diatom (*Thalassiosira pseudonana*) [9,10,16]. For both *Arabidopsis* sets and for honeybee, data from each individual array showed a nearly identical shift in the cytosine enrichment of the top 5% of probes (data not shown). In contrast, the exact shifts were different among each of the four *Chlamydomonas* arrays, although the overall trend was similar. Despite being synthesized with the same maskless technology, the human, rice (*Oryza sativa*), and fruit fly (*Drosophila melanogaster*) arrays did not share the nucleotide bias among their high signal probes, and yeast (*Saccharomyces cerevisiae*) data showed substantially lower bias [3,5,7,17]. Thus it was unclear if the C-rich correlation with high signal was either due to a genome-specific cause, or a difference in laboratory techniques. Also notable was the relative absence of nucleotide bias in the sea urchin (*Strongylocentrotus purpuratus*) tiling data, the only genome examined with 50-mer microarray probes [11]. Several possible explanations were considered for these observations (see Supplementary Discussion).

2.2. Oligomeric motif analysis

To determine if other factors contributed to the C-rich signal bias, it was examined whether a random distribution of high C-content across a probe was sufficient to be associated with high signal, or if clusters of cytosines were more important. Moreover, we were curious if other neighboring nucleotides had signal enhancing effects. To investigate this phenomenon, we devised a method that reflects the representation of short oligomeric sequences among the probes with the strongest array signals. For a given N-mer a score was assigned as the median signal intensity of all probes containing the N-mer sequence, divided by the median signal of all probes on the array. Scores were calculated for all permutations of N-mers of 2–9 bases in length, and the N-mers were ranked by score for further examination.

$$\underset{N2\rightarrow 9}{\text{N-mer Score}}\ (S) = \frac{m_{\text{N-mer}}}{m_{\text{all probes}}},$$

where m is the median signal intensity.

This analysis was performed on oligonucleotide array data from multiple genomes, and representative results are shown in Fig. 2.

It became apparent that the score for the top oligomeric sequence of a given length approached a logarithmic increase in proportion to the length of the N-mer (Fig. 2A). Furthermore, a relatively small proportion of the N-mers were associated with large changes in signal intensity. In contrast, the bottom range of scores remained relatively unaffected by changes in N-mer length. The scores of longer N-mers should approach a limit when additional neighboring nucleotides cease to be important. However, the observation of dramatically rising scores coupled to increasing complexity indicated that an unexpectedly long sequence specificity was likely to be found among the top-scoring motifs. Further analysis of the 9-mer sequences showed a striking C-rich bias in nucleotide composition among the highest scoring 1% of the motifs (Fig. 2B). This was consistent with our earlier observation of increased C-content among high signal probes.