Design and use of external controls in DNA microarray experiments

Supplementary information for

"In control: systematic assessment of microarray performance"

Harm van Bakel and Frank C.P. Holstege

Published in EMBO reports in 2004

External RNA controls can be an important tool in determining microarray performance. External controls are synthetically produced RNA molecules with corresponding probes on the microarray. Control RNAs can be added to biological samples of interest in defined amounts. When spiked in at different levels between the two channels of a cDNA microarray, they can be used to assess the ability to detect differentially expressed genes (see accompanying paper). Other uses include normalization, sensitivity testing (Dorris *et al.*, 2002; Girke *et al.*, 2000; Hughes *et al.*, 2001; Lockhart *et al.*, 1996; Ramakrishnan *et al.*, 2002; Relogio *et al.*, 2002), background estimation (Dorris *et al.*, 2002; Ramakrishnan *et al.*, 2002), and the comparison of labelling strategies (Badiee *et al.*, 2003). Detailed recommendations on how to design and handle external controls and corresponding microarray probes are described below.

RNA control design

The most important requirement for external controls is that they are representative of the endogenous mRNAs with regards to length and sequence characteristics. In addition, the external controls need to be designed so that cross-hybridization towards and from endogenous transcripts is avoided. The present availability of whole genome sequences enables this criteria to be more easily met (Bozdech *et al.*, 2003). Preferably, a low risk of cross-hybridization will still be present such that the controls are representative of endogenous mRNAs. This allows the technology to be optimized for avoiding cross-hybridization artifacts. For these and for other reasons presented below, it is imperative that multiple controls are used.

The external controls should also be representative with regard to the presence and size of polyA tails. For example, in labelling protocols using modified (d) UTP, this nucleotide will be incorporated in the antisense cDNA or cRNA, in particular at the position corresponding to the polyA tail. If the polyA tails are significantly shorter, control RNAs will be labelled less efficiently and will also be less susceptible to fluorescent quenching of closely incorporated dyes (Ramdas *et al.*, 2001; t Hoen *et al.*, 2003) in the polyA tail. This might lead to aberrant reporting of control RNA levels compared to endogenous mRNAs.

One way of potentially avoiding artifacts from different length polyA tails is to use a two nucleoside-anchored polyT primer in the cDNA synthesis step (F.C.P.H, unpublished results).

Probe design

Probes representing the control RNAs must be typical of all the probes on the microarray. This dictates that they should be designed with the same criteria of melting temperature, sequence content, secondary structure etc.(Bozdech *et al.*, 2003), as the other probes. Cross-hybridization can occur in two directions. Endogenous mRNAs can cross-hybridize with control probes and control RNAs can cross-hybridize with gene probes. In order to achieve representation and avoid cross-hybridization, all the probes on a microarray should be designed bearing the control RNA sequences in mind, i.e. by treating them for the purpose of design, as part of the genome or collection of genes in question. The two directions of cross-hybridization can be tested experimentally by comparing spiking to not spiking. Cross-hybridization will manifest itself as gene probes reacting to the addition of control RNAs and significant signals on control probes in the absence of spikes. An additional interesting test is to determine whether there is a good correlation between the amount of spike control added and fluorescent intensity (main article, figure 3). This is more of a test of overall microarray performance than external control behavior. However, outliers might indicate aberrant controls or probes in this test.

To ensure that controls are representative, it is important to have the control probes manufactured and processed in parallel with the other probes. This is because probes can suffer from several different batch-dependent artifacts during production and processing. One example is batch-dependent amounts of contaminating material in probe preparations that gives rise to increased intensity in the Cy3 channel, possibly explaining one of the systematic sources of dye bias described recently (Martinez *et al.*, 2003; Raghavachari *et al.*, 2003).

Vector requirements

Generation of control RNA material is a straightforward process that makes use of *in vitro* transcription by bacteriophage T7 or T3 RNA polymerase. Generally T7 results in higher yields, but both are in the 50-100 μg range from single reactions driven by 0.5 μg of template DNA. It is essential that the templates are completely, but not excessively, linearized by restriction digestion so that correct transcripts are produced. For generating external control probes for inclusion on PCR-product based cDNA arrays, it is logical to use the RNA polymerase template plasmids as templates for the PCR reactions too.

In our experience such probes do not work well, perhaps because polyA tails and too much vector sequence are included when using common PCR primers. One solution is to generate separate plasmids for generating PCR products. These PCR probes should include only the external control RNA sequence, flanked by the same primers used for the rest of the cDNA clone collection. The polyA tail should be excluded from the probe, assuming that this is also the case for the cDNA collection.

Handling external controls

For routine use of external controls, several measures need to be taken to avoid batch- and storage-dependent artifacts that confound the analysis of microarray performance over longer periods of time. The high yield of RNA produced by T7 RNA polymerase readily lends itself to creating stocks for many thousands of experiments. External control RNAs at high concentrations (>1mg/ml) can be stored for at least three years at -80°C, without noticeable loss of signal. Lower, more useful concentration mixes should be made from these stocks and also stored at -80°C for shorter periods (6 months). High and low concentration stocks should be discarded after three rounds of freeze-thawing to counter RNA fragmentation and should be aliquoted bearing this in mind.

The final, useful concentration mixes of external controls should be assembled so that the volume added is not prone to significant pipetting errors. For most combinations of humans and liquid-handling devices, the safest approach is that the control should be no less and preferably more than 5µl, regardless of the volume of the sample. Addition of lower amounts such as 1 µl is possible, but high precision and accuracy is then restricted to certain combinations of liquid-handling devices and people. The replication of experiments will in most cases reveal pipetting inaccuracy that can perhaps be reduced by the introduction of automated systems.

Incorporation on microarrays

To represent the entire mRNA population, it is essential that multiple external controls are used. In addition, multiple probes for each control should be distributed all over the array. This allows local hybridization artifacts to be monitored. For example, assessing the homogeneity of hybridizations by different automated hybridization stations can be readily achieved by evaluating how closely the differently located probes return similar values, i.e. how tight the bunching is of the control groups in figure 2. Supplemental figure 1 shows an example of a microarray subgrid incorporating external controls for self-spotted arrays.

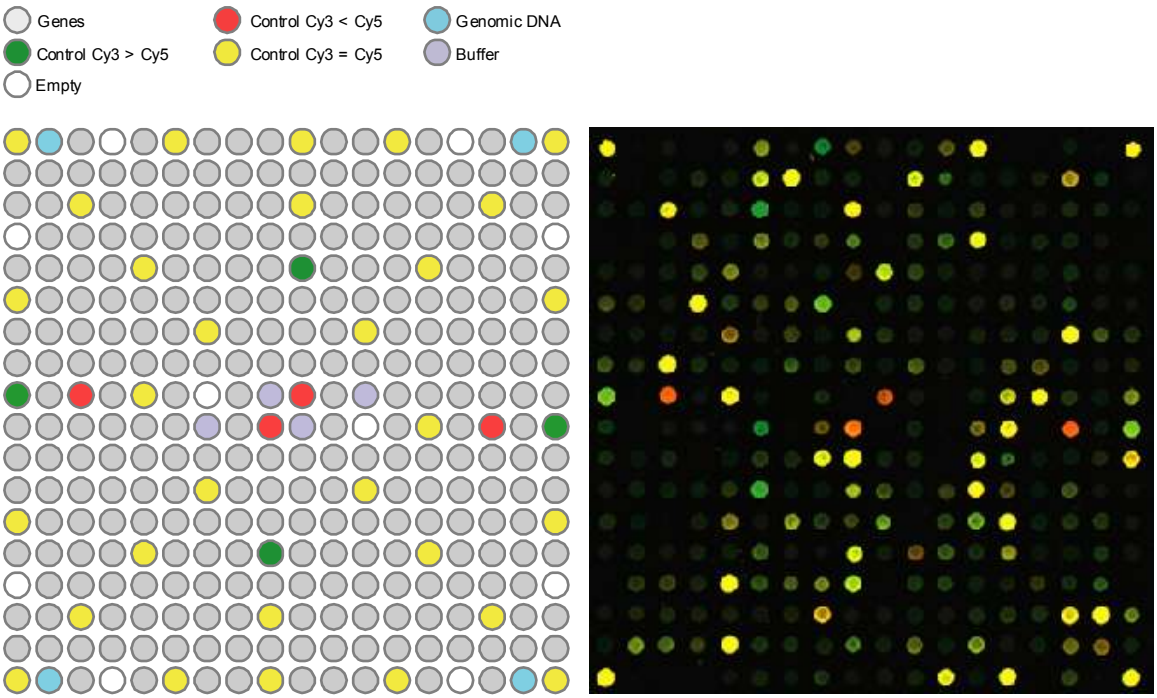

Figure 1: Incorporating external controls

Example of a microarray subgrid layout with external controls. (A) Schematic overview of array subgrid. Ratio and normalisation controls are placed in a radial pattern to obtain unbiased measurements across the entire subgrid. Each control is spotted at least twice. Additional controls include checks for spotting and hybridisation artefacts (empty, genomic DNA and buffer spots). (B) Self versus self hybridisation with external controls spiked according to 1A.

Including each probe at least once per subgrid on spotted arrays, also allows the performance of different spotting pins to be monitored. This is important because the robots that are currently employed by most facilities to manufacture arrays use up to 48 different spotting pins to deposit DNA. Differences between pins can lead to pin bias (Yang *et al.*, 2002). If required, local non-linear normalization on each subgrid can also be performed using just control probes (van de Peppel *et al.*, 2003). This requires sufficient numbers of external controls and/or replicate probes to be included on each subgrid. Finally, probes representing high level spiked controls can be positioned as anchor spots in all subgrid corners for easier grid alignment during image quantification.

References

Badiee, A., Eiken, H.G., Steen, V.M. and Lovlie, R. (2003). Evaluation of five different cDNA labeling methods for microarrays using spike controls. *BMC Biotechnol.* **3:** 23.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B. and DeRisi, J.L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol.* **4:** R9.

Dorris, D.R., *et al.* (2002). A highly reproducible, linear, and automated sample preparation method for DNA microarrays. *Genome Res.* **12:** 976-984.

Girke, T., Todd, J., Ruuska, S., White, J., Benning, C. and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol.* **124:** 1570-1581.

Hughes, T.R., *et al.* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol.* **19:** 342-347.

Lockhart, D.J., *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.* **14:** 1675-1680.

Martinez, M.J., *et al.* (2003). Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays. *Nucleic Acids Res.* **31:** e18.

Raghavachari, N., Bao, Y.P., Li, G., Xie, X. and Muller, U.R. (2003). Reduction of autofluorescence on DNA microarrays and slide surfaces by treatment with sodium borohydride. *Anal Biochem.* **312:** 101-105.

Ramakrishnan, R., *et al.* (2002). An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Res.* **30:** e30.

Ramdas, L., *et al.* (2001). Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biol.* **2:** RESEARCH0047.

Relogio, A., Schwager, C., Richter, A., Ansorge, W. and Valcarcel, J. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* **30:** e51.

t Hoen, P.A., de Kort, F., van Ommen, G.J. and den Dunnen, J.T. (2003). Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Res.* **31:** e20.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. and Holstege, F.C. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep.* **4:** 387-393.

Yang, Y.H., *et al.* (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30:** e15.