BMC Genomics

BioMed Central

Research article

Open Access

# Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data

Rachel Lyne†[1,2], Gavin Burns†[1], Juan Mata[1], Chris J Penkett[1], Gabriella Rustici[1], Dongrong Chen[1], Cordelia Langford[1], David Vetrie[1] and Jürg Bähler*[1]

Address: [1]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, U.K and [2]Present address: Department of Genetics, University of Cambridge, Cambridge CB2 3EH, U.K

Email: Rachel Lyne - rachel@flymine.org; Gavin Burns - gpb@sanger.ac.uk; Juan Mata - jm6@sanger.ac.uk; Chris J Penkett - cjp@sanger.ac.uk; Gabriella Rustici - gr2@sanger.ac.uk; Dongrong Chen - dcm@sanger.ac.uk; Cordelia Langford - cfl@sanger.ac.uk; David Vetrie - vt1@sanger.ac.uk; Jürg Bähler* - jurg@sanger.ac.uk

* Corresponding author †Equal contributors

Published: 10 July 2003

*BMC Genomics* 2003, **4**:27

Received: 20 March 2003
Accepted: 10 July 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/27

© 2003 Lyne et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

## Abstract

**Background:** The genome of the fission yeast *Schizosaccharomyces pombe* has recently been sequenced, setting the stage for the post-genomic era of this increasingly popular model organism. We have built fission yeast microarrays, optimised protocols to improve array performance, and carried out experiments to assess various characteristics of microarrays.

**Results:** We designed PCR primers to amplify specific probes (180–500 bp) for all known and predicted fission yeast genes, which are printed in duplicate onto separate regions of glass slides together with control elements (~13,000 spots/slide). Fluorescence signal intensities depended on the size and intragenic position of the array elements, whereas the signal ratios were largely independent of element properties. Only the coding strand is covalently linked to the slides, and our array elements can discriminate transcriptional direction. The microarrays can distinguish sequences with up to 70% identity, above which cross-hybridisation contributes to the signal intensity. We tested the accuracy of signal ratios and measured the reproducibility of array data caused by biological and technical factors. Because the technical variability is lower, it is best to use samples prepared from independent biological experiments to obtain repeated measurements with swapping of fluorochromes to prevent dye bias. We also developed a script that discards unreliable data and performs a normalization to correct spatial artefacts.

**Conclusions:** This paper provides data for several microarray properties that are rarely measured. The results define critical parameters for microarray design and experiments and provide a framework to optimise and interpret array data. Our arrays give reproducible and accurate expression ratios with high sensitivity. The scripts for primer design and initial data processing as well as primer sequences and detailed protocols are available from our website.

## Background

DNA microarrays are currently one of the most powerful and widespread technologies for functional genomics, allowing the study of genome-wide gene expression and

other global DNA-dependent processes (for reviews see [1–3]). Glass microarray slides contain thousands of nucleic acid features that can be interrogated in parallel. In the popular two-colour assay, fluorescently labelled samples prepared from RNA of two different cell populations are co-hybridised onto the microarray to measure relative gene expression levels [4,5]. In the following, we will refer to the known features spotted on the microarray as 'array elements' and to the labelled cDNA hybridised to the microarray as 'samples'.

Although becoming increasingly routine, DNA microarrays are still not a 'plug-and-play' technology, requiring substantial optimisation for reliable performance. A wide range of factors can affect data quality, and it is important to understand various parameters involved [6–9]. This includes conditions that lead to unwanted changes in gene expression before RNA extraction, such as variations in the environment (e.g., temperature shocks) or in the genotype (e.g., auxotrophic markers). Other parameters, including array and experimental design, protocols and data processing procedures, can affect array data independently of biological processes. To use microarrays to their full potential, it also helps to know the performance characteristics of a microarray platform. Important properties are data reproducibility within and between arrays, effects of dye biases and other artefacts in the data structure, as well as the accuracy, specificity and sensitivity of signal intensity measurements.

The fission yeast *Schizosaccharomyces pombe* is a popular model organism whose genome has been fully sequenced [10]. The genome is well annotated and contains somewhat less than 5000 predicted genes [11,12]. Fission yeast has a low-complexity genome and similar experimental advantages as the budding yeast *Saccharomyces cerevisiae*, which has been widely used to pioneer functional genomics approaches (reviewed in [13,14]). Experimental conditions for yeast can be tightly controlled, and it is straightforward to study homogeneous populations of cells and to combine findings from global studies with genetic data. Fission and budding yeasts are only distantly related and separated >1000 million years ago according to recent estimates [15]. *S. pombe* therefore provides a valuable complementary model system, and it should be insightful to compare and contrast global datasets obtained in these two unicellular eukaryotes.

We have recently reported genome-wide expression profiles during sexual differentiation and stress responses in fission yeast [16,17]. Here we describe the design of the fission yeast microarrays together with experimental procedures and data evaluation pipeline. Various properties and performance characteristics of our microarray system were measured, which help to understand the nature and limitations of array data.

## Results and Discussion

### *Effects of array element properties on fluorescence signals and ratios*

We have built a DNA microarray containing elements for all the known and predicted open reading frames (ORFs) of the fission yeast genome (see Methods). For each ORF, we amplified 180–500 bp of exon sequence by PCR. To test the effect of array element size on fluorescence signal intensities, PCR products of a wider size range were used for some genes (80–1500 bp). Although the signals tended to increase with increasing element size, there was only a ~2-fold difference between the lowest and highest signal intensities (Figure 1A). Array elements larger than ~500 bp did not lead to increased signals, similar to what has been reported before [18,19]. Importantly, the signal ratios were independent of array element size, and they were much less variable than the signal intensities, differing by only a few percents (Figure 1A).

The position of an array element within a gene also affected the signal intensities. Figure 1B shows that tiles of equal length selected closer to the 3' end of genes tend to give higher signals than those closer to the 5' end, and there was a ~3-fold difference between lowest and highest signal intensities. This reflects the efficiency of reverse transcription, which is primed from the polyA-tail of the mRNAs. Again, the signal ratios showed little variation and were independent of signal intensities, suggesting that both fluorescence dyes are equally incorporated during labelling and independently of array element positions (Figure 1B). Because signal intensities can affect reproducibility, we maximized signals by selecting array elements that were less than 2.5 kbp from the gene ends (see Methods).

For each array element, we performed two rounds of PCR, using gene-specific primer pairs for the first round and gene-specific reverse primers in combination with a universal forward primer containing a 5'-aminolink for the second round. This allowed covalent linkage of the coding strand to the modified glass slides, thus providing a single-stranded array element that is specific for transcriptional direction. Available data strongly suggest that our array elements are indeed strand-specific. For example, *mek1* transcripts did not give any microarray signals above background in timecourse experiments of cells undergoing meiosis and sporulation [16], although *mek1* is induced during meiosis [20]; we later realized that the primers for *mek1* had been designed the wrong way round and that our microarray had included only the anti-sense strand of this gene. Moreover, the *rec7* transcript showed a different gene expression profile during the meiotic

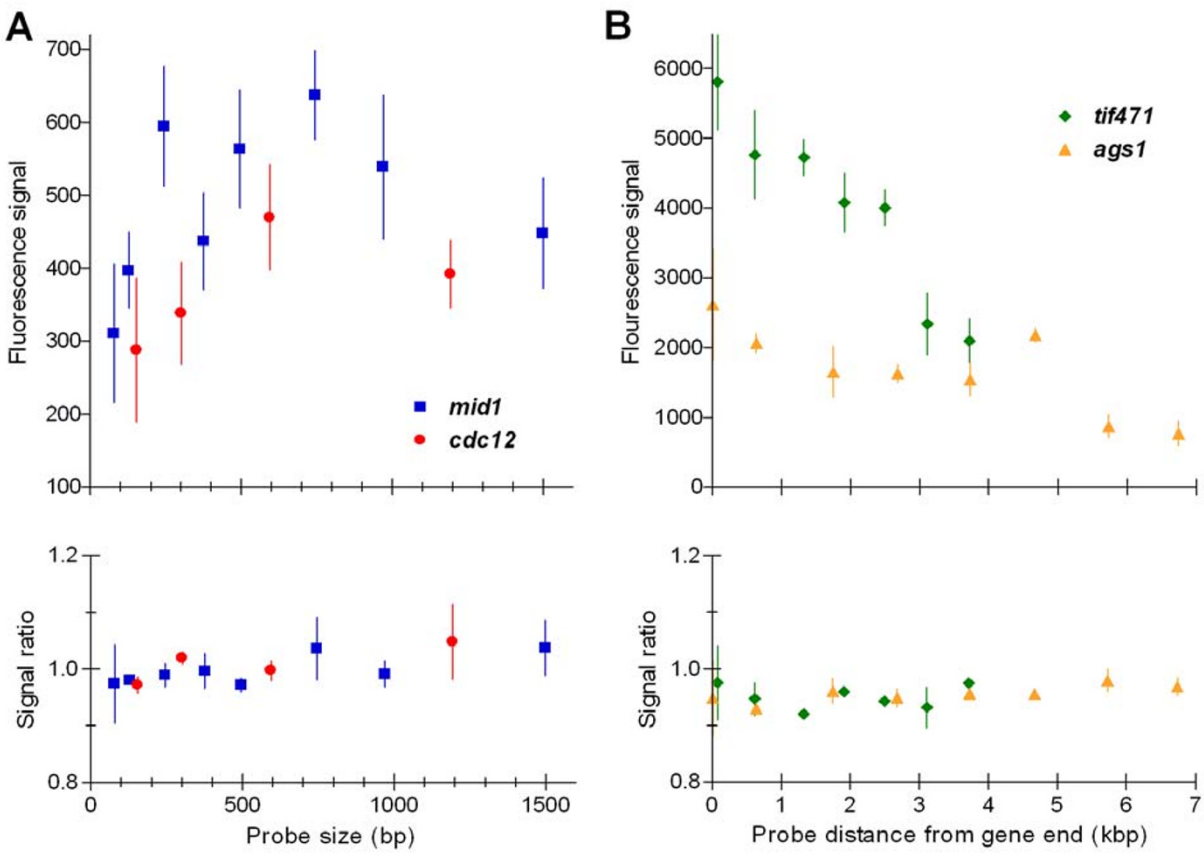


**Figure 1**
Effects of array element size and position on fluorescence signals and ratios. **(A)** PCR products of varying sizes (80–1500 bp) were used as array elements for two genes (*mid1* and *cdc12*). In all cases, the 3' ends of the array elements were kept constant (~50 bp upstream of the stop codon). Top: fluorescence signals (local background subtracted) relative to array element size; the means and standard deviations of eight signal measurements are shown (two self-self experiments with two replicate measurements of both Cy3 and Cy5 each). Bottom: normalized ratios of signals (Cy5/Cy3) relative to array element size; the means and standard deviations of four measurements are shown (two self-self experiments with two replicate measurements each). **(B)** PCR products from varying positions within two genes (*ags1* and *tif471*) were used as array elements. In all cases, the sizes of array elements were similar (~500 bp). Top: fluorescence signals (local background subtracted) relative to array element position (measured as distance of 3'ends of elements to stop codon); the means and standard deviations of eight signal measurements are shown (two self-self experiments with two replicate measurements of both Cy3 and Cy5 each). Bottom: normalized ratios of signals (Cy5/Cy3) relative to array element position; the means and standard deviations of four measurements are shown (two self-self experiments with two replicate measurements each).

timecourse experiments compared to the overlapping *tos1*, *tos2*, and *tos3* that are transcribed from the opposite strand [16,21]. The strand specificity was also evident with genes for non-coding RNAs where we selected array elements for both orientations (Table 1). In all cases, the elements containing the coding strands produced strong signals, whereas the signals from the non-coding strands were too close to the background signals to pass our cutoff criteria. We conclude that only the amino-modified strands bind significantly to the slides. This allowed the design of strand-specific array elements that can discriminate transcriptional directions and that minimize the interference of complementary strands during hybridization. Accordingly, single-stranded array elements give higher signals than double-stranded elements of the same size and sequence on Codelink slides (D.V., unpublished observations).

**Table 1: Strand-specific array elements report direction of transcription**

| Gene [Reference] | Array element direction [a] | Mean signal +/- SD [b] |
|---|---|---|
| *meiRNA* [42] | forward | 2790 +/- 850 |
| | reverse | 30 +/- 17 |
| *meu3* [37] | forward | 25440 +/- 12210 |
| | reverse | 40 +/- 37 |
| *meu11* [37] | forward | 1040 +/- 250 |
| | reverse | 8 +/- 5 |
| *meu19* [37] | forward | 35960 +/- 13470 |
| | reverse | 70 +/- 40 |
| *meu20* [37] | forward | 5100 +/-1050 |
| | reverse | 30 +/- 20 |

[a] forward: coding strand contains 5'-amino modification; reverse: non-coding strand contains 5'-amino modification. [b] Mean and standard deviation of six fluorescence signals (3 experiments with two replicate measurements each) from meiotic *pat1* timecourse at 5 hr timepoint [16].

### *Initial data processing and normalization*

We developed a script for data pre-processing and normalization (for details, see Methods). To filter out unreliable data, we only use signals from those spots with more than 50% of the pixels greater than two standard deviations (SD) above local background signal in both channels. Data from spots showing more than 95% of the pixels greater than 2 SD above local background in one channel are retained, even if the other channel does not pass the normal cutoff. This prevents the elimination of data from genes that are not or very weakly expressed in one condition, but show reliable expression in the other condition. Although the absolute ratio values from such spots will not be accurate, it is valuable to identify these genes as they are expected to be clearly differentially expressed. The script also provides a quality control report indicating the number and percentages of discarded spots as well as data of genes with low correlation between replicate spots.

DNA microarray data are based on signal ratios, and the relative fluorescence intensities between the scanned channels must be normalized to adjust for systematic biases such as differences in RNA levels, dye incorporation, and detection efficiencies. To visualize the global structure of microarray data, it is popular to use ratio-intensity plots (also referred to as MA plots) [22,23]. These plots can reveal signal intensity-dependent biases in ratio measurements caused by non-linear effects of fluorescence dyes at extreme signals. Figure 2A shows that the log signal ratios deviate from zero in the low signal range. A proposed way to correct for such effects is by Lowess normalization [8,24], although the actual reasons for the imbalances that Lowess corrects are not well understood [25]. After filtering out data from weak signals as described above, the remaining signals do not appear to show intensity-dependent biases (blue spots in Figure 2A). Moreover, results were very similar either with or without Lowess normalization using GeneSpring (data not shown). We therefore do not apply this type of normalization to our data.

The positions of the spots on the array, unlike the signal intensities, did have a pronounced effect on the signal ratios in many of our arrays (Figure 3A). The pattern and strength of these spatial effects varies from slide to slide. For the purpose of illustration, we show an extreme case of this phenomenon, but the effects we normally observe are much milder. These spatial artifacts could be caused by uneven slide surface or differences in hybridization conditions across the array, by the scanning process, or by a combination of all (unpublished observations; see also [26]). The spatial effects varied continuously from region to region and did not correlate with sub-grids of the array that are printed with different spotting pins. It is therefore unlikely that the observed irregularities are caused by the spotting pins, although pin variations can lead to local differences in signal ratios in some cases [24]. These spatial artifacts cannot be corrected with a global normalization, and our script therefore uses a local normalization scheme. For each spot on the array, we apply a sliding window of neighboring spots (see Methods for details). Assuming that the median ratio of all spots showing measurable signals within the sliding window should be 1, the script calculates a normalization factor for the signal ratio of the central spot. The array elements are ordered randomly on the array with regard to chromosome position, minimizing the probability that our normalization scheme masks any biological effects caused by genome rearrangements such as deletions or duplications.

The application of this local normalization scheme leads to ratios that are well balanced over the entire measurable signal range (Figure 2B,2C) and that no longer show any position-dependent effects (Figure 3B,3C). To further test

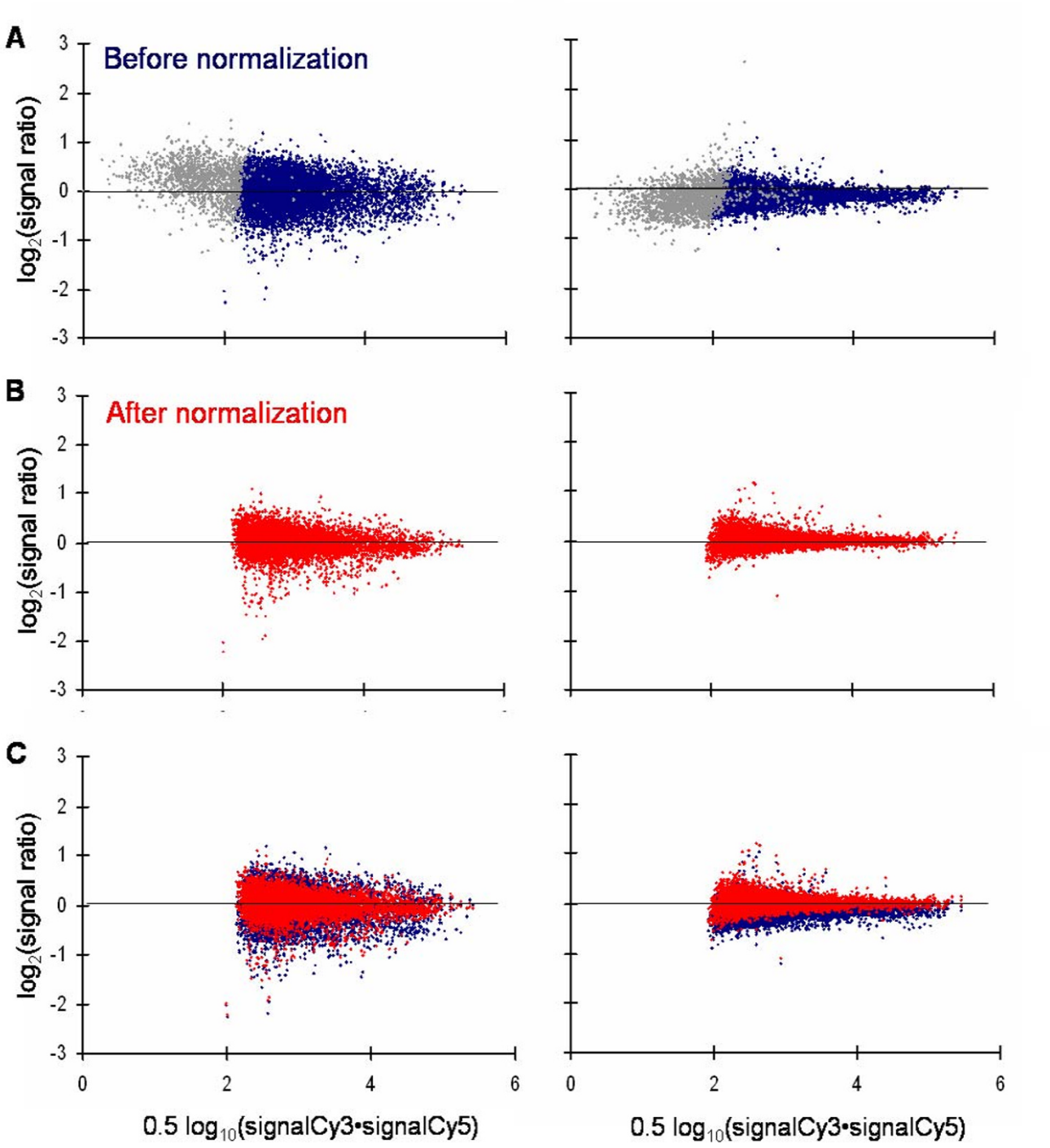


**Figure 2**
Ratio-intensity plots before and after normalization. **(A)** log-transformed signal ratios (Cy3/Cy5) are plotted against the log-transformed products of signal intensities. Grey spots: discarded data that were filtered out by initial data processing (see text for details); blue spots: data used for further evaluation. Left side: temperature experiment with sample from cells grown at 25°C labelled with Cy5 and sample from cells grown at 30°C labelled with Cy3. Right side: identical sample labelled with Cy3 and Cy5 and hybridised on same array ('self-self' hybridisation): all signal ratios are expected to be 1, and the absence of differential gene expression is reflected by a tighter distribution of the spots. The number of the blue spots that were retained for data evaluation is 9161 (left) and 8560 (right). **(B)** As in (A) after normalization of the data using our local normalization scheme. **(C)** Overlay of spots before (blue) and after (red) normalization.

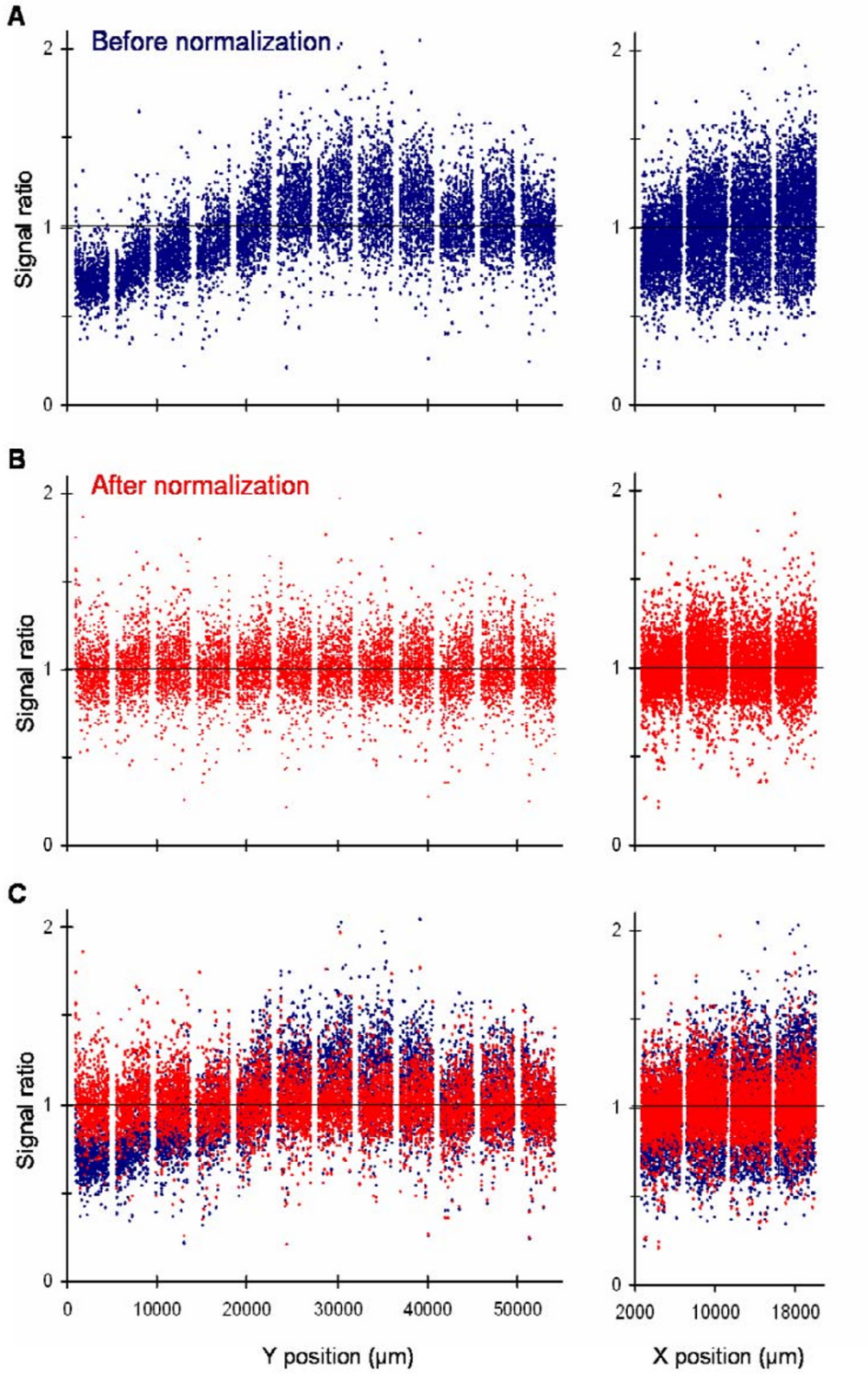


**Figure 3**
Correction of spatial artifacts by local normalization. **(A)** Distribution of signal ratios along the Y- (left) and X-axis (right) of the microarray slide before normalization. The data are from the same array as in Figure 2 (left side). Only spots giving usable data are shown. The groups of spots separated by small gaps reflect the 12 × 4 sub-grids of the array, each printed with a different spotting pin. **(B)** Distribution of signal ratios as in (A) after local normalization of the data. **(C)** Overlay of the data from (A) and (B).

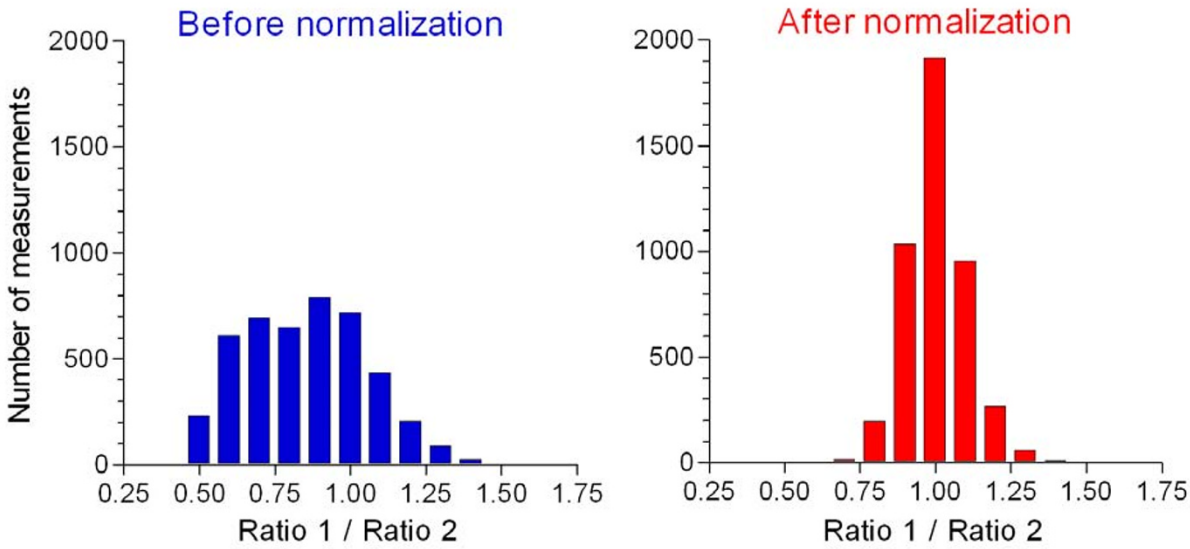


**Figure 4**
Replicate data are more similar to each other after local normalization. The ratios of signal ratios from corresponding pairs of replicate spots on the same array were determined for all array elements (Ratio 1 / Ratio 2). Increased agreement between replicate measurements leads to ratios of signal ratios closer to 1. The data are from the same array as in Figure 2 (left side) and Figure 3. Left histogram: replicate data distribution before normalization; right histogram: replicate data distribution after normalization.

whether our normalization scheme improves the data, we compared data from replicate spots on an array before and after normalization. Indeed, replicate data are in better agreement with each other after local normalization, consistent with higher data quality (Figure 4). While our normalization performs well in most biological conditions, the assumption that the median of ratios is 1 may not be valid in some cases (e.g., in quiescent cells where much of transcription is shut down). If there are large differences in gene expression between the samples, it should be preferable to use external controls for normalization.

#### *Sensitivity, specificity, and accuracy of microarray data*

The fission yeast microarrays contain a range of control spots including elements for five bacterial genes (see Methods). Spots with these bacterial genes do not pick up any measurable signals when hybridized with fission yeast cDNA (negative controls). Spiking of known quantities of the corresponding bacterial mRNAs into the labelling mix allows to estimate the linear range and sensitivity of our arrays (Figure 5). The signal readout is linear over a wide range of at least 3–3000 pg amount of transcript. We can easily measure fluorescence signals from transcript amounts as low as 3 pg (in a complex sample of 20 μg of total RNA). This should allow detection of one mRNA molecule in a population of at least 400,000, thus allowing the measurement of very weakly expressed genes, given the low-complexity fission yeast genome. The number of genes that produce measurable signals depends on biological conditions and on array quality. For exponentially growing cells, we typically measure 80–90% of all genes. Some 350 genes did not give measurable signals in any of six self-self experiments of cells vegetatively growing in rich (4 experiments) or minimal medium (2 experiments), suggesting that these genes are not or very lowly expressed in vegetative cells. There was no significant difference in the number of measurable genes between cells grown in rich vs minimal medium. Although we cannot directly correlate these data with the number of genes actually being expressed, it appears that the majority of genes show at least a basal level of expression during exponential vegetative growth, given that negative control genes do not give any signals above background. In humans fewer genes appear to be expressed in a given cell line [27], although it can be as high as 80% for some tissue culture samples [22].

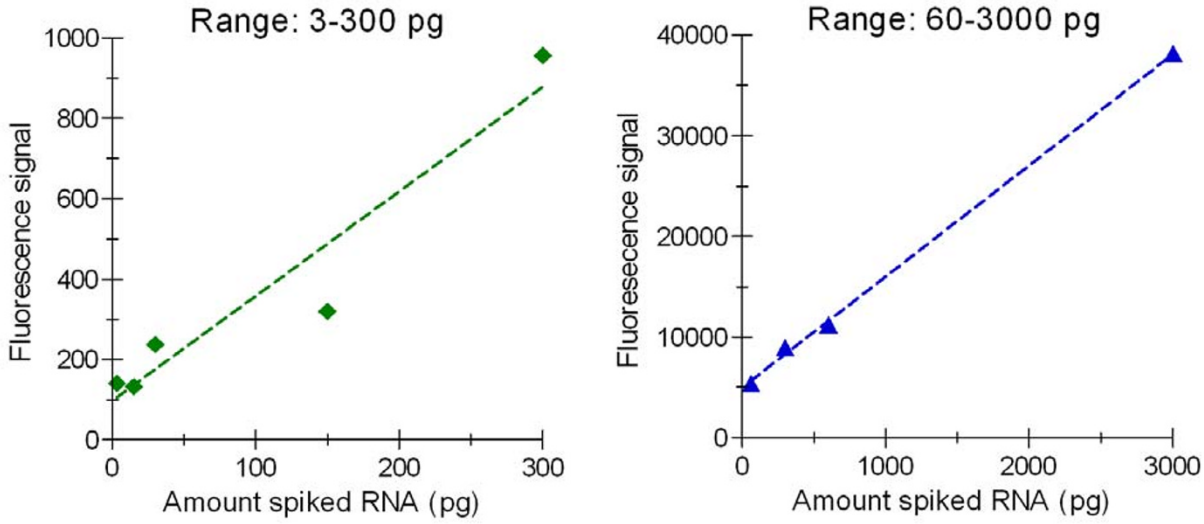


**Figure 5**
Linear readout range and detection limit of spiked RNA samples. 20 μg of *S. pombe* total RNA was spiked with five *Bacillus subtilis* mRNAs at various concentrations. Left side: lysA (3 pg), pheB (15 pg), thrB (30 pg), dapB (150 pg), and trpC (300 pg); right side: lysA (60 pg), pheB (300 pg), thrB (600 pg), and dapB (3000 pg). The median fluorescence intensities above local background of the *B. subtilis* control spots (determined from ~100 spots/transcript distributed evenly across the array) were plotted as a function of transcript concentration.

The arrays also contain elements for a range of budding yeast genes that show varying degrees of DNA sequence identity to specific fission yeast genes (22 array elements ranging from 30–79% identity to *S. pombe* elements). This allows to estimate the level of sequence diversity required between two genes to give specific signals. Figure 6 shows hybridization data of the 13 array elements with the highest similarity to *S. pombe* genes. Both the absolute 'unspecific' signal intensities and the percentages of 'unspecific' relative to 'specific' signals are shown, because highly expressed genes are expected to result in higher absolute 'unspecific' signal intensities. Cross-hybridization becomes apparent with array element sequence identities higher than ~70% under our optimized hybridization and washing conditions. Array elements with lower than 70% identity had very weak signals without exception that would not pass our cutoff criteria. The *S. pombe* array elements were selected to be less than ~70% identical to other regions in the genome wherever possible (see Methods). We therefore expect that there will be only few skewed data due to cross-hybridization.

We also determined the accuracy of the signal ratios measured with our arrays using the *S. cerevisiae* elements. To this end, we spiked *S. cerevisiae* RNA in various amounts into the labelling reactions and compared expected with measured signal ratios after microarray hybridization (Table 2). The median signal ratios were close to the expected ratios in all cases and for a wide range of signal intensities; the least accurate measurements were ~40% lower than expected. *S. cerevisiae* array elements with more than 70% homology to *S. pombe* genes gave lower than expected ratios (probably due to competitive hybridization from *S. pombe* samples) and were not included in the data of Table 2. We conclude that our arrays give accurate data for signal ratios within the range tested.

### *Reproducibility of array data*

Information from repeatedly spotted array elements and repeatedly hybridised experiments help to assess data quality and reliability. Reproducibility can be measured at different levels: 1) duplicate spots within an array, 2) hybridisation of an identical sample to different arrays (technical repeats), and 3) hybridisation of independent samples from two identical experiments (biological repeats). As measures for reproducibility, we have calculated the standard deviation (SD) and coefficient of variation (CV) between repeatedly measured signal ratios (Table 3). Generally, the microarrays give highly reproducible results with typically no or very few genes

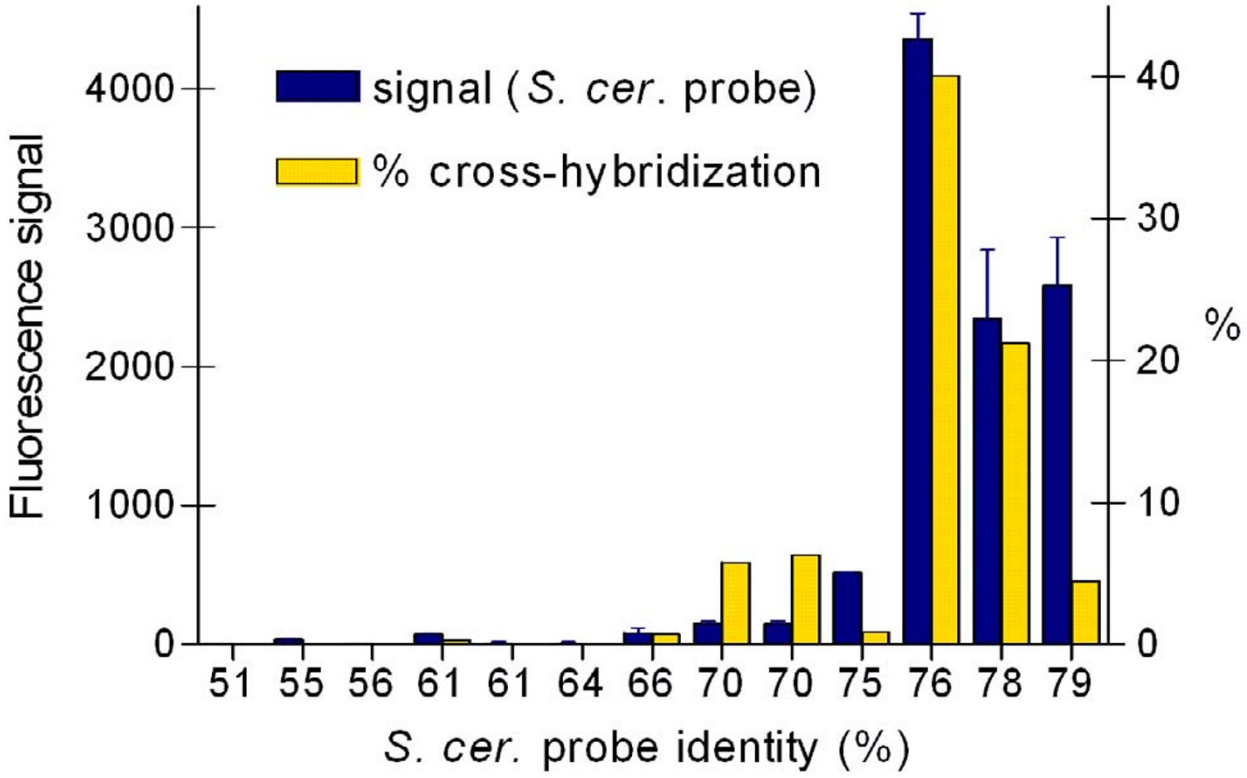


**Figure 6**
Unspecific hybridisation to similar array elements. PCR products from *S. cerevisiae* genes with various sequence similarities to *S. pombe* genes were used as array elements and hybridised with *S. pombe* samples. Hybridisation data are shown for 13 such genes showing 51%-79% identity to *S. pombe* genes across their entire lengths of ~200 bp. Blue bars: fluorescence signals (local background subtracted) picked up by the *S. cerevisiae* array elements; the means and standard deviations of four signal measurements are shown (one self-self experiment with two replicate measurements of both Cy3 and Cy5 each). Yellow bars: relative amount (in percentages) of 'unspecific' signals from *S. cerevisiae* array elements compared to 'specific' signals from the corresponding *S. pombe* array elements. Array elements for the following *S. cerevisiae* genes were used (increasing similarity; corresponding *S. pombe* genes in parentheses): *HDA1* (SPAC8C9.06c); *RPL18A* (*rpl18-1*); *CDC2* (*cdc6*); *CDC19* (*pyk1*); *RPL18A* (*rpl18-2*); URA7 (SPAC10F6.03c); *RPL27A* (*rpl27-2*); *HOG1* (*sty1*); *YPT1* (*ypt2*); *ACT1* (*act1*); *HTA1* (*hta1*); *HTB1* (*htb1*); *ACT1* (*act1*).

**Table 2: Accuracy of signal ratios determined by spiking of *S. cerevisiae* RNA**

| Spiked ratios [a] | Measured median ratios (range)[b] | Range of signals[c] |
|---|---|---|
| 1:2 | 1.9 (1.8–2.1) | 490/270 – 62,080/32,060 |
| 1:5 | 5.5 (4.3–7.2) | 420/90 – 58,320/10,990 |
| 1:10 | 9.7 (5.8–11.1) | 380/40 – 35,890/5620 |
| 1:20 | 19.6 (12.2–23.6) | 340/20 – 56,900/3110 |

[a] Total *S. cerevisiae* RNA was used in the following amounts: 3 μg:6 μg; 2 μg:10 μg; 1 μg:10 μg; 0.5 μg:10 μg. For normalization, 10 μg of total *S. pombe* RNA was included in each labelling reaction. [b] Medians of 30 signal ratios (15 *S. cerevisiae* array elements with two replicate measurements each). The range indicates the lowest and highest signal ratios measured. [c] The *S. cerevisiae* array elements produce a wide range of signals; the lowest and largest signal pairs used to determine signal ratios are shown.

**Table 3: Reproducibility of array data**

| Measurement | Mean SD (Range) [a] | CV (Range) [b] |
|---|---|---|
| Within array replicates [c] | 0.04 (0.03–0.06) | 4.4% (3.1–6.2%) |
| Technical repeats [d] | 0.04 (0.02–0.06) | 4.5% (2.5–6.3%) |
| Biological repeats [e] | 0.07 (0.05–0.10) | 6.4% (4.9–8.1%) |

[a] Standard deviation (SD) of signal ratios were calculated for each measurable pair of array elements, and the mean SDs of all repeated measurements are indicated. The range indicates the lowest and highest mean SDs obtained from several pairs of comparisons. [b] CV: Coefficient of variation: [(SD of signal ratios × 100)/mean of signal ratios] was calculated for each measurable pair of array elements, and the mean CVs of all repeated measurements are indicated. The range indicates the lowest and highest mean CVs obtained from several pairs of comparisons. [c] Determined from 10 arrays of experiments used in this study (4 self-self, 1 temperature, 4 media, and 1 harvesting experiment; see Methods). [d] Determined from 7 pairs of arrays from this study (4 self-self, 4 media) and from [17](oxidative stress); for each pair identical RNA samples were used that were labelled independently and with reverse colors for the two hybridisations. Within array replicate data were averaged before analysis. [e] Determined from 9 pairs of arrays from this study (4 self-self, 4 media) and from [17](oxidative stress); for each pair RNA samples from identical but independent biological experiments were used that were labelled with reverse colors for the two hybridisations. Within array replicate data were averaged before analysis.

showing greater than 2-fold differences between different measurements (Figure 7A). Technical and biological repeats were compared after averaging data from duplicate spots within one array, which produces more accurate measurements and reduces variability. For straightforward biological experiments (such as the comparison of cells logarithmically growing in different media, Figure 7A), biological repeats gave similar reproducibility to technical repeats. However, the variability of biological repeats tended to be higher than for technical repeats in many experiments where biological conditions could not be as tightly controlled (Table 3). Because the technical variability is consistently low in our arrays, we always use samples prepared from independent biological experiments to obtain repeated measurements on different arrays. This is the most effective use of microarrays as the biological variability, which is the source of the greatest noise in most cases, is included in the repeated measurements (see also [6]).

To check for systematic biases in fluorescent dye incorporation, we performed a series of self-self hybridisations, i.e., an identical sample was labelled with both the Cy3 and Cy5 fluorochrome and hybridised on the same array. A typical scatter plot of such a hybridisation is shown in Figure 7B. The great majority of genes appeared to be similarly labelled with both dyes, and amongst the spots with good data (blue) no signal intensities were greater than 2-fold different in the two channels. The average SD of signal ratios from self-self experiments was 0.08. However, if self-self experiments were repeated using the same sample, a number of genes appeared as significantly differentially expressed. For example, using significance analysis of microarrays (SAM) allowing one false positive [28], no genes appeared as differentially expressed with up to three technical repeats, but 108 genes were identified as 'differentially expressed' with four technical repeats. If the signal ratios from two of these self-self experiments were inverted, no differentially expressed genes were found by SAM. This indicates that the 108 'significant' genes do not reflect a stochastic process but a systematic bias for some genes, most likely in dye incorporation. This bias is too subtle to be evident with few repeats, but became statistically significant if the experiment was repeated more than three times. To prevent this dye bias, we routinely swap the dyes during labelling for repeated hybridisations.

## Conclusions

We present a complete description of our microarray platform for fission yeast together with various data on array performance and properties that are rarely reported. This helps to compare our microarrays with other platforms, and it provides a framework to put array data into context and understand their potential and limits. We also report straightforward and reliable experimental procedures together with a data evaluation pipeline optimised for the fission yeast microarrays. This results in accurate, sensitive, and highly reproducible data, allowing reliable measurements of expression ratios of the great majority of all fission yeast genes. The reported procedures and resources should also be useful for other microarray systems. To get the best data out of a limited number of arrays, we recommend to use samples from independent biological experiments and to swap the fluorescent dyes for repeated hybridisations. All experimental protocols, primer sequences, and the scripts for primer design and initial data processing are available from our website [29].

## Methods

### *Microarray construction: primer design, PCR reactions, and arraying*

We generated each array element by polymerase chain reaction (PCR) using gene-specific primer pairs (GenSet)

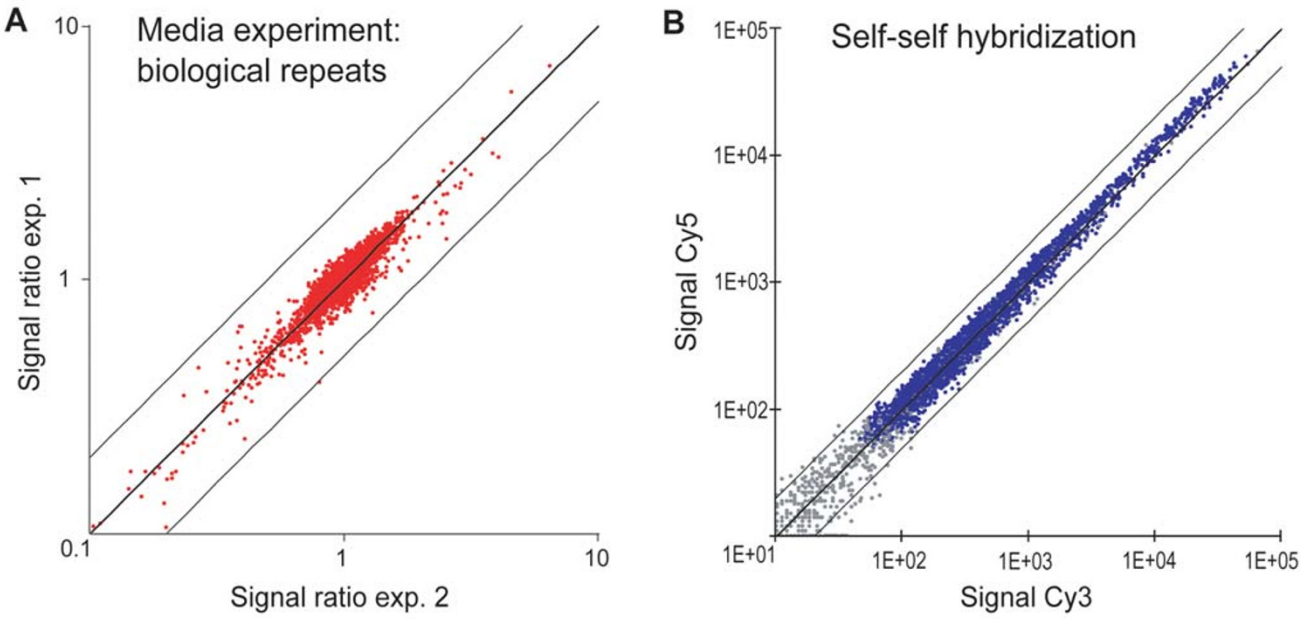


**Figure 7**
Reproducibility of signal ratios and intensities. **(A)** Scatter plot showing the reproducibility between two biological repeats of an experiment where cells grown in minimal or rich media were directly compared to each other. The plot represents 4245 genes that gave measurable data in both experiments. The CV for the repeated experiment shown here is 5%. Just one gene shows an ~2-fold difference in ratios between the two experiments (just outside the outer lines). **(B)** Autocorrelation plot showing the distribution of Cy5 and Cy3 signal intensities from a single self-self experiment. Median signal intensities minus median local background intensities were determined and signals from replicate spots were averaged. Grey spots: data from 790 spots that were flagged 'absent' during analysis or initial data processing (see Methods). Blue spots: data from 4260 spots that were retained for evaluation. All the signal intensities from the blue spots are <2-fold different from each other (within outer lines).

selected for each of the predicted and known ORFs in the annotated *S. pombe* genome sequence [10,11]. We wrote a Perl script (available at our website: [29]) to batch process EMBL format files for exon selection and primer processing. PRIMER3 [30,31] was used to determine primer sequences matching defined criterions. The majority of primers were 18–22 bp long with melting temperatures between 58–62°C and GC contents between 40–60%. Primers were selected such that the resulting amplicons were 180–500 bp long and contained 100% exon sequence, and the reverse primers were positioned <2500 bp upstream of the stop codon. All the forward primers had an additional 8 bp universal sequence at their 5' end (5'-TGACCATG-3'), which is not included in above parameters. All primer and amplicon sequences were blasted against the *S. pombe* genome. Only primers and amplicons that showed no significant similarity to other sequences in the genome were used (i.e., primers with a blast score of <70 and amplicons with a blast score of <400, the latter corresponding to less than ~70% sequence identity). For ~50 genes, we amplified up to 150 bp of 3'- or 5'-untranslated regions to obtain more specific array elements. In a few cases of highly similar genes, we had to use less specific array elements (blast score of <1000 with other sequences in the genome); this affected ~140 genes, including many ribosomal protein and transposon-related genes.

In addition to the predicted ORFs, we amplified fragments of the 11 mitochondrial genes, 19 pseudogenes, various RNA genes (a few genes for ribosomal RNA, tRNAs, and snRNAs as well as 68 other larger genes for 'miscellaneous RNAs' [32]), 114 very hypothetical ORFs, 33 large introns, as well as centromeric repeats and ars elements. The latest microarrays contain elements for 5269 different genes and other genomic features of fission yeast. Some genes are represented by two or more different array elements. We also designed array elements from 22 *S. cerevisiae* genes showing varying degrees of similarity to *S. pombe* genes to control for cross-hybridization. The arrays also contain elements for several widely used markers and epitope tagging sequences: Kan-MX, GFP, GST, Myc, and

3HA [33]; TAP [34]; and Pk [35]. A detailed file containing all the primer sequences and parameters is available from our website [29]. PCR products of five genes from the prokaryote *Bacillus subtilis* were used as control elements on the array (lysA, pheB, dapB, thrB, and trpC). These can be used as positive controls by spiking in a 'cocktail' of the corresponding bacterial mRNAs in known quantities (for details on control genes and preparation of mRNA 'cocktails' by *in vitro* transcription, see [36]).

PCR reactions were performed in 96-well plates (Costar) using a Tetrad thermocycler (MJ Research). For each array element, two rounds of PCR reactions were performed. For the first PCR reaction, we used gene-specific primer pairs, with forward primers containing an additional universal sequence (see above). As a template, we used genomic DNA prepared with a simple glass bead protocol [33]. To amplify array elements from genes containing only small exons (<250 bp), we used pools of cDNA libraries as a template ([37,38]; pREP3X: constructed by B. Edgar and C. Norbury; Clontech). PCR products from the first round were used as templates for the second round of PCR reactions, together with gene-specific reverse primers and a universal forward primer containing a 5'-amino modification (5'-GCTGAACAGCTATGACCATG-3'; Oswel). Details of the PCR reaction mixes and cycling parameters are available from our website [29]. All PCR products were checked for single strong bands of expected sizes on 2.5% agarose 1x TBE slab gels. Typically, the failure rate was <3%. Failed PCR reactions were repeated, and new primer sequences were ordered in cases where PCR reactions failed repeatedly. At the time of writing, array elements for all predicted genes had been successfully amplified. The gene-specific primer pairs together with the two sequential and independent PCR reactions make it highly unlikely that array elements are assigned to wrong genes.

Spotting buffer was added to the PCR products at a final concentration of 250 mM sodium phosphate pH 8.5, 0.00025% Sarkosyl, followed by spin filtration using 96-well filtration plates (Millipore). The filtered array elements with spotting buffer (15 μl total volume) were then re-arrayed into 384 well plates (Genetix), snap frozen on dry ice, and stored at -70°C. These array elements were printed without any further purification onto activated amine-binding slides (Codelink, Amersham) using a BioRobotics TAS arrayer with a 48-pin tool. All array elements are printed in duplicate onto each slide (~13,000 spots/slide). The replicate spots are printed in separate halves of the slides and with different spotting pins to obtain two measurements that are as independent as possible [6], and to prevent local depletion of the sample and minimize the chance of losing both measurements of a gene due to local hybridisation problems (unpublished observations). One array of each batch was quality control tested by hybridization. Array elements were dried completely in a vacuum concentrator and stored at -70°C in sealed plates between print rounds. Before printing, array elements were reconstituted by addition of HPLC water (BDH) and left to dissolve o/n at 4°C. Details of the arraying and post-processing procedures are available from the website of the Microarray Facility at the Sanger Institute [36].

### *RNA isolation from fission yeast*

We used the *S. pombe* wild-type strain 972 $h^-$ for all experiments [39]. Standard media and growth conditions were used [40], and cells were harvested from liquid cultures at mid-exponential phase ($OD_{600}$ 0.1–0.4), unless stated otherwise. For the spike-in experiment (Table 2), *S. cerevisiae* cells (strain AB1380) were grown in YPD medium to $OD_{600}$ 0.3, and RNA was extracted as described below for *S. pombe* cells.

Cells were harvested either by mild centrifugation (2 min, 800 rcf), and the pellet was snap frozen in liquid nitrogen after discarding the supernatant, or by rapid filtration (Millipore), and the filters were snap frozen in liquid nitrogen after transfer into a 50 ml tube. To see whether these two methods of cell harvesting affect gene expression, we used a microarray to directly compare RNA samples obtained after cell filtration or centrifugation of the same culture grown in EMM medium. The data obtained from the two samples were very similar to each other (SD of signal ratios: 0.08), and only two mitochondrial genes were 2-fold different between the samples. We conclude that the two methods of cell harvesting that we routinely use do not lead to significant differences in gene expression.

Total RNA was isolated from *S. pombe* cells using a hot phenol method followed by phenol-chloroform extractions, precipitation, and purification using Qiagen RNeasy columns. (We had also experimented with isolating mRNA before labelling, and only a few genes give different results compared to total RNA. Because mRNA isolation requires much larger cell samples and potentially introduces biases, we routinely use total RNA for labelling.) RNA quality was determined by gel electrophoresis and spectrophotometry. A detailed protocol is available from our website [29].

### *Sample labelling and microarray hybridisation*

To generate fluorescently labelled samples for microarray hybridisation, we used a direct labelling protocol. 10–20 μg of total RNA was reverse transcribed into cDNA with Superscript enzyme (GibcoBRL) and an oligo-$dT_{17}$ primer in the presence of Cy3- or Cy5-dCTP (PerkinElmer). We have also experimented with a mix of random nonamer

and oligo(dT) primers for labelling; although this will lead to amplification of non-coding RNAs, it does not give increased background, and significantly improves the signal intensities of most spots. Only a few genes are differentially labelled when comparing the two priming methods. One advantage of using a random primer is that mRNAs without or with short polyA tails will also be represented in the hybridisation. The labelled cDNAs were purified using AutoSeq G-50 columns (Amersham) and precipitation. Hybridization was performed at 49°C in a buffer containing 48% formamide using LifterSlips (Erie Scientific) and a hybridisation oven with humid chamber (Boekel Scientific). Slides were washed at room temperature and stored in the dark for scanning. A detailed protocol for labelling, hybridisation, and slide washing is available from our website [29].

### *Data acquisition, processing, normalization, and evaluation*

Microarrays were scanned using a GenePix 4000 B laser scanner, and fluorescence signals were analysed using GenePix Pro software (Axon Instruments). Array images that did not pass minimal quality thresholds were not used (median signal-to-background >3; median signal-to-noise >5; mean of median background signal <200). Technically flawed spots were removed either automatically by the GenePix software or through manual investigation of the array images, and such spots were flagged as 'absent' in the GenePix results files.

For subsequent data processing and normalization, we developed a Perl script that uses GenePix results files as input (script available from or website [29]). This script discards data from spots with failed or faulty PCR products by masking them 'absent'. Data from spots with low array element concentration (as judged by PCR product staining on gel) or PCR products where the reverse primer is located 2500–3500 bp from the gene end are flagged 'marginal'. All genes on the array are also represented by at least one good array element, and 'marginal' data from sub-optimal array elements are only used as a backup if other data from a given gene are not available. The script also applies cut-off criteria to discard data from weak signals: spots with <50% of pixels >2 SD above median local background signal in one or both channels are flagged 'absent', unless one channel shows >95% of the pixels >2 SD above local background. The SD was calculated using only the lower 55% of the pixel intensities (called SD2 in GenePix Pro), as this measure is less susceptible to being skewed by bright pixels. The script provides a quality control report showing the numbers and percentages of spots discarded during the various steps of the data analysis pipeline as well as data of replicate spots with signal ratios >2-fold different from each other.

The script also performs a local normalization using a sliding square window of spots surrounding each spot. A user-defined minimum number of spots is chosen to be used with which to normalize over (default is 400). The window size default is 16 spots. This means the square contains 33 × 33 spots (1089) surrounding central spots, 33 × 17 spots (561) surrounding spots at the edge of the array, and 17 × 17 spots (289) surrounding spots in a corner of the array. Only spots that are flagged 'present' are used for the normalization. Hence, using a window of 16 means that sometimes, especially for spots close to the corners of the array, less than 400 spots may actually be used for the normalization. In cases where the block size chosen is small, the window size is increased up to a user-defined maximum window size (default is 24) so that at least 600 total spots are in the square. This means the block size used with this window change is larger than may be necessary to optimise the chances of having 400 'present' spots to use for normalization. This is a heuristic to make the algorithm faster for the majority of spots, since counting the number of 'present' spots in the initial square uses a relatively large amount of computational time. If, during normalisation, the number of spots is still found to be less than 400, the window is increased further until the maximum window size is reached. In these cases, the spots that do use less than 400 spots for normalisation are reported in the output log file. The script then calculates a normalization factor such that the median signal ratio of all measurable spots within the square equals 1, and this factor is then used to scale the signal ratio for the central spot. The signal ratios used for normalization correspond to the median of all pixel-by-pixel ratios of signals minus median local background for each pixel of a given spot (called 'median of ratios' in GenePix Pro). This measures ratios more reliably and is less affected by unspecific signals than the 'ratio of medians' (see also [41]). In the rare cases where the 'median of ratios' was zero, the 'ratio of medians' was used instead for data evaluation. Finally, the script averages the normalized data from all replicate spots that produced measurable signal ratios of the same genomic element. These mean normalized ratios were then used for downstream data evaluation and mining using GeneSpring (Silicon Genetics) and SAM [28].

### *Microarray experiments used in this study*

Self-self experiments were performed with RNA isolated from exponentially growing cells, followed by labelling identical samples with both Cy3 and Cy5 fluorochromes and hybridising on the same array (six experiments in total). Self-self experiments were used for data in Figure 1, Figure 2 (right), Figure 5, Figure 6, Figure 7B, Table 2, and Table 3. For experiments showing differential gene expression, we compared samples from cells growing at 25°C vs 30°C (one experiment; used for Figure 2 [left], Figures

3,4,5, and Table 3), samples from cells growing in full vs minimal medium (four experiments; used in Figure 7A and Table 3), as well as samples from cells harvested by centrifugation vs filtration (one experiment; used in Table 3). Some data were acquired from previously published experiments, including samples from meiotic vs vegetative cells ([16]; Table 1) and samples from oxidatively stressed vs unstressed cells ([17]; Table 3).

## Authors' contributions

RL wrote the computer scripts for primer design and initial data processing, gave informatics support, and contributed to data evaluation. GB amplified the array elements, developed optimised protocols, performed most of the microarray experiments, and contributed to data evaluation. JM developed the scheme for local normalization and participated in data evaluation. CJP refined the computer scripts and participated in data management and organization. GR and DC contributed to protocol optimisation and some experiments. CL and DV helped to build and print the arrays and provided initial protocols. JB conceived the study, participated in its design and coordination, and drafted the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We thank Val Wood for help with genome information, the Microarray Facility at the Sanger Institute for arraying, Aengus Stewart and Mike Lyne for help with primer design, and Hiroshi Nojima for a meiotic cDNA library.

This work was funded by Cancer Research UK.

## References

1. Brown PO and Botstein D: **Exploring the new world of the genome with DNA microarrays** *Nat Genet* 1999, **Suppl 21:**33-37.
2. Lockhart D and Winzeler E: **Genomics, gene expression and DNA arrays** *Nature* 2000, **405:**827-836.
3. Young R: **Biomedical discovery with DNA arrays** *Cell* 2000, **102:**9-16.
4. Schena M, Shalon D, Davis RW and Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray** *Science* 1995, **270:**467-470.
5. Shalon D, Smith SJ and Brown PO: **A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization** *Genome Res* 1996, **6:**639-645.
6. Churchill GA: **Fundamentals of experimental design for cDNA microarrays** *Nat Genet* 2002, **32(Suppl 2):**490-495.
7. Holloway AJ, Van Laar RK, Tothill RW and Bowtell DD: **Options available-from start to finish-for obtaining data from DNA microarrays II** *Nat Genet* 2002, **32(Suppl 2):**481-489.
8. Quackenbush J: **Microarray data normalization and transformation** *Nat Genet* 2002, **32(Suppl 2):**496-501.
9. Yang YH and Speed T: **Design issues for cDNA microarray experiments** *Nat Rev Genet* 2002, **3:**579-588.
10. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J and Baker S *et al.*: **The genome sequence of *Schizosaccharomyces pombe*** *Nature* 2002, **415:**871-880.
11. **Schizosaccharomyces pombe GeneDB** [http://www.genedb.org/genedb/pombe/index.jsp]
12. Wood V and Bähler J: **How to get the best from fission yeast genome data** *Comp Funct Genom* 2002, **3:**282-288.
13. Delneri D, Branca FL and Oliver SG: **Towards a truly intergrative biology through the functional genomics of yeast** *Curr Opin Biotechnol* 2001, **12:**87-91.
14. Kumar A and Snyder M: **Emerging technologies in yeast genomics** *Nat Rev Genet* 2001, **2:**302-312.
15. Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL and Hedges SB: **Molecular evidence for the early colonization of land by fungi and plants** *Science* 2001, **293:**1129-1133.
16. Mata J, Lyne R, Burns G and Bähler J: **The transcriptional program of meiosis and sporulation in fission yeast** *Nat Genet* 2002, **32:**143-147.
17. Chen D, Toone WM, Mata J, Lyne R, Burns G, Kivinen K, Brazma A, Jones N and Bähler J: **Global transcriptional responses of fission yeast to environmental stress** *Mol Biol Cell* 2003, **14:**214-229.
18. Stillman BA and Tonkinson JL: **Expression microarray hybridization kinetics depend on length of the immobilized DNA but are independent of immobilization substrate** *Anal Biochem* 2001, **295:**149-157.
19. Franssen-van Hal NL, Vorst O, Kramer E, Hall RD and Keijer J: **Factors influencing cDNA microarray hybridization on silylated glass slides** *Anal Biochem* 2002, **308:**5-17.
20. Perez-Hidalgo L, Moreno S and San-Segundo PA: **Regulation of meiotic progression by the meiosis-specific checkpoint kinase Mek1 in fission yeast** *J Cell Sci* 2003, **116:**259-271.
21. Molnar M, Parisi S, Kakihara Y, Nojima H, Yamamoto A, Hiraoka Y, Bozsik A, Sipiczki M and Kohli J: **Characterization of *rec7*, an early recombination gene in *Schizosaccharomyces pombe*.** *Genetics* 2001, **157:**519-532.
22. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J and Li J *et al.*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays** *Genome Biol* 2002, **3:**research0062.
23. Yang YH and Speed T: **Representing and evaluating slide data** In *DNA Microarrays* Edited by: *Bowtell D, Sambrook J. New York: Cold Spring Harbor Laboratory Press*; 2002:544-551.
24. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP: **Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation** *Nucleic Acids Res* 2002, **30:**e15.
25. Verdnik D, Handran S and Pickett S: **Key considerations for accurate microarray scanning and image analysis** In: *DNA Array Image Analysis: Nuts & Bolts* Edited by: *Kamberova G, Shishir S. DNA Press LLC*; 2002:83-98.
26. Colantuoni C, Henry G, Zeger S and Pevsner J: **Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts** *Biotechniques* 2002, **32:**1316-1320.
27. Snijders AM, Meijer GA, Brakenhoff RH, van den Brule AJC and van Diest PJ: **Microarray techniques in pathology: tool or toy?** *Mol Pathol* 2000, **53:**289-294.
28. Tusher VG, Tibshirani R and Chu G: **Significance analysis of microarrays applied to the ionizing radiation response** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
29. **Bähler Lab Website: Pombe Functional Genomics** [http://www.sanger.ac.uk/PostGenomics/S_pombe/]
30. Rozen S and Skaletsky HJ: **Primer3 on the WWW for general users and for biologist programmers** In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* Edited by: *Krawetz S, Misener S, Totowa NJ. Humana Press*; 2000:365-386.
31. **Primer3 Code** [http://www-genome.wi.mit.edu/genome_software/other/primer3.html]
32. Watanabe T, Miyashita K, Saito TT, Nabeshima K and Nojima H: **Abundant Poly(A)-bearing RNAs that lack open reading frames in *Schizosacharomyces pombe*.** *DNA Research* 2002, **9:**209-215.
33. Bähler J, Wu JQ, Longtine MS, Shah NG, McKenzie A III, Steever AB, Wach A, Philippsen P and Pringle JR: **Heterologous modules for efficient and versatile PCR-based gene targeting in *Schizosaccharomyces pombe*** *Yeast* 1998, **14:**943-951.
34. Tasto JJ, Carnahan RH, McDonald WH and Gould KL: **Vectors and gene targeting modules for tandem affinity purification in *Schizosaccharomyces pombe*** *Yeast* 2001, **18:**657-662.
35. Craven RA, Griffiths DJ, Sheldrick KS, Randall RE, Hagan IM and Carr AM: **Vectors for the expression of tagged proteins in *Schizosaccharomyces pombe*** *Gene* 1998, **221:**59-68.

36. **The Microarray Facility at the Sanger Institute** [http://www.sanger.ac.uk/Projects/Microarrays/]
37. Watanabe T, Miyashita K, Saito TT, Yoneki T, Kakihara Y, Nabeshima K, Kishi YA, Shimoda C and Nojima H: **Comprehensive isolation of meiosis-specific genes identifies novel proteins and unusual non-coding transcripts in *Schizosaccharomyces pombe*** *Nucleic Acids Res* 2001, **29:**2327-2337.
38. Fikes JD, Becker DM, Winston F and Guarente L: **Striking conservation of TFIID in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*** *Nature* 1990, **346:**291-294.
39. Leupold U: **Genetical methods for *Schizosaccharomyces pombe*** *Meth Cell Physiol* 1970, **4:**169-177.
40. **Fission Yeast Handbook** [http://www.sanger.ac.uk/PostGenomics/S_pombe/docs/nurse_lab_manual.pdf]
41. Brody JP, Williams BA, Wold BJ and Quake SR: **Significance and statistical errors in the analysis of DNA microarray data** *Proc Natl Acad Sci USA* 2002, **99:**12975-12978.
42. Watanabe Y and Yamamoto M: ***S. pombe mei2*$^{+}$ encodes an RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA*** *Cell* 1994, **78:**487-498.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

**BioMed**central