The impact of quantitative microarray optimization on gene expression analysis

Peter Sykacek[*†‡] David P. Kreil[*†] Lisa A. Meadows[§]
Richard P. Auburn[§] Bettina Fischer[§] Steven Russell[§]
Gos Micklem[§¶]

Revision 2nd August 2010

— Supplement —

Enquiries: Peter.Sykacek@boku.ac.at

[*] Joint first authorship.
[†] Chair of Bioinformatics, Department of Biotechnology,
Boku University, Vienna,
A-1190 Muthgasse 18, Austria
[‡] Corresponding author
[§] Cambridge Systems Biology Centre, University of Cambridge,
Tennis Court Road, Cambridge, CB2 1QR, U.K.,
& Department of Genetics, University of Cambridge,
Downing Street, Cambridge CB2 3EH, U.K.
[¶] Cambridge Computational Biology Institute, Dept of Applied Mathematics
and Theoretical Physics, University of Cambridge, Wilberforce Road,
Cambridge CB3 0WA, U.K.

Required Declarations

The supplementary material which this document references is for confidential preview only. Please treat all data accordingly.

We cannot take responsibility for the availability of material referenced by links to external sites. Please advise us if you encounter any problems.

Main paper abstract

Background:

With the growing availability of entire genome sequences, an increasing number of scientists can exploit oligonucleotide microarrays for genome-scale expression studies. While probe-design is a major research area, relatively little work has been reported on the optimization of microarray protocols.

Results:

As shown in this study, suboptimal conditions can have considerable impact on biologically relevant observations. For example, deviation from the optimal temperature by a centigrade lead to a loss of 44% of differentially expressed genes identified. While genes from thousands of Gene Ontology categories were affected, transcription factors and other low-copy-number regulators were disproportionately lost. Calibrated protocols are thus required in order to take full advantage of the large dynamic range of microarrays.

For an objective optimization of protocols we introduce an approach that maximizes the amount of information obtained *per* experiment. A comparison of two typical samples is sufficient for this calibration. We ensure, however, that optimization results are independent of the samples and the specific measures used for calibration. Both simulations and spike-in experiments confirm an unbiased determination of *generally* optimal experimental conditions.

Conclusions:

Well calibrated hybridization conditions are thus easily achieved and necessary for the efficient detection of differential expression. They are essential for the sensitive profiling of low-copy-number molecules. This is particularly critical for studies of transcription factor expression, or the inference and

study of regulatory networks.

This online supplement is available from http://bioinf.boku.ac.at/pub/optMA2010/. Data and scripts are also available from that archive. Please note that the supplement and auxiliary materials are under Copyright

Contents

Chapter S-1

Using this archive

S-1.1 Viewing the Supplement and material referenced

This document is provided in PDF format (*cf.* Section S-1.2). Auxiliary information is referenced by HTTP URLs (Hyper Text Transfer Protocol – Universal Resource Locations). If you view this document in a stand-alone browser, *e. g.*, Acrobat Reader, clicking on a link should open a new browser window showing the content to which the link refers.

If you are viewing this document through a plug-in, your browser may loose the original page context when following a link, so when you go back to this document, you might return to the title page. In such a case you may want to save this document to a local disk, and then view it in a stand-alone PDF browser, like Acrobat Reader.

S-1.2 Description of file formats

File formats used in this archive include the following:

- American Standard Code for Information Interchange (ASCII) is used in data files, pre-formatted text for reports, and program code / script files. Columns in data files are typically TAB delimited.
  This format is the simplest and should cause the least problems.

- Adobe Portable Document Format (PDF) for typeset material. This supplement is made available in PDF.

  There are free viewer programs available for this format. To obtain such a viewer, please visit, for example:

  – Ghostscript, Ghostview and GSview from the Computer Sciences Department at the University of Wisconsin-Madison, USA,

  – Adobe Acrobat Reader from Adobe Inc., USA.

  Many browser programs for the World Wide Web can run so-called plug-ins for viewing PDF content.

- Adobe PostScript (PS) for typeset material. To obtain free tools for viewing and printing, please visit, for example, Ghostscript, Ghostview and GSview from the Computer Sciences Department at the University of Wisconsin-Madison, USA. These files are provided for convenience only, and are usually the best format for printing.

- `bzip2` compressed files. Large files (particularly text) may be compressed with `bzip2` for efficiency. Free utilities to unpack such files are available from `http://www.bzip.org/`.

- Microsoft Excel spreadsheet files (XLS) for certain charts and tables.

- Grace plotting tool files (AGR). The original files containing the data to produce some graphs and figures. Although they are text files, you may prefer to view them in the freely available Grace plotting program.

- Various graphics file formats. Typical formats include JPEG, which is a lossy compression format well suited for photos with smooth gradients, and TIFF, which is a particularly flexible format, supporting

both lossy and non-lossy compression schemes (TIFF-FAQ). For viewing or converting many graphics file formats, free tools are available (GraphicsMagick, ImageMagick).

- ZIP archives. Larger collections of files are provided in compressed archives. Free utilities to unpack these archives are available from the Info-ZIP group. Users of the Microsoft Windows system may wish to use WinZIP.

Chapter S-2

Results

S-2.1 Cross-hybridization tables (ΔG)

Systematic all-*vs*-all calculations allowed a comprehensive assessment of predicted cross-hybridization potential. Here, we first consider extremely conservative thresholds supporting the full observed dynamic range of microarray signals (*1*), $R = 10^6$, with a contamination ratio $r < 1\%$, which in the worst case scenario requires that $\Delta G_{\text{diff}} > 12.6\,\text{kcal/mol}$ at $T_{\text{hyb}}^{(\text{eff})} = 70°\text{C}$; see Methods, Eq. (1), in the main manuscript. In order to observe cross-hybridization at this threshold, the contaminating sequence needs to be in 10^6-fold excess of the target sequence and we must be able to detect intensity deviations as low as 1% of the total signal.

A less extreme but still conservative threshold supports the typical dynamic range of expression intensities seen in a single microarray image scan, $R = 10^3$, with $r < 10\%$ contamination, in the worst case scenario, if $\Delta G_{\text{diff}} > 6.3\,\text{kcal/mol}$. Results were filtered accordingly with each set of thresholds, predicting cross-hybridization potential for 2.5% of probes, with worst-case cross-hybridization detectable in single-scan microarray quantification for only 2.2% of probes. For comparison, predictions for a commercial library were 5.7% and 5.2% of probes, respectively.

In summary, calculations indicated little and weak cross-hybridization for the examined probe design. A detailed description of the design including cross-hybridization matrices by probe and by target is provided below.

Besides the raw unfiltered output, tables were compiled at the two cut-offs described above. Table variants are available to show cross-hybridization by target and by probe for both the recently designed INDAC FL002 platform (Kreil *et al.*, submitted). There are table variants showing the probes' target IDs ('CG'-number) and variants showing the internally used unique probe IDs. Each row of the tables starts with the target/probe ID and is followed by match groups. Each match group contains the match type ('ok', or 'X' for cross-hybridization), the match ID, and the match ΔG. All fields are tab delimited.

S-2.2 Calibration of $T_{\mathbf{hyb}}^{(\mathbf{eff})}$

S-2.2.1 Results overview

Results from the complementary assays described in the paper are collected in Table 1 of the main manuscript. The below figure provides a graphical summary.

S-2.2.2 Calculation of generalization accuracy by cross-validation

Complementing these results we provide predictive accuracy estimates based on six-fold cross-validation, in which both channels of one of the six individual slides in turn were used as test samples. All transcripts are used individually to predict whether a channel measurement indicates a male or female sample. For every probe a probit link generalized linear model (GLM) was fitted for the discrimination of male and female samples. The resulting receiver operating characteristic (ROC) curves, estimates of the mutual in-

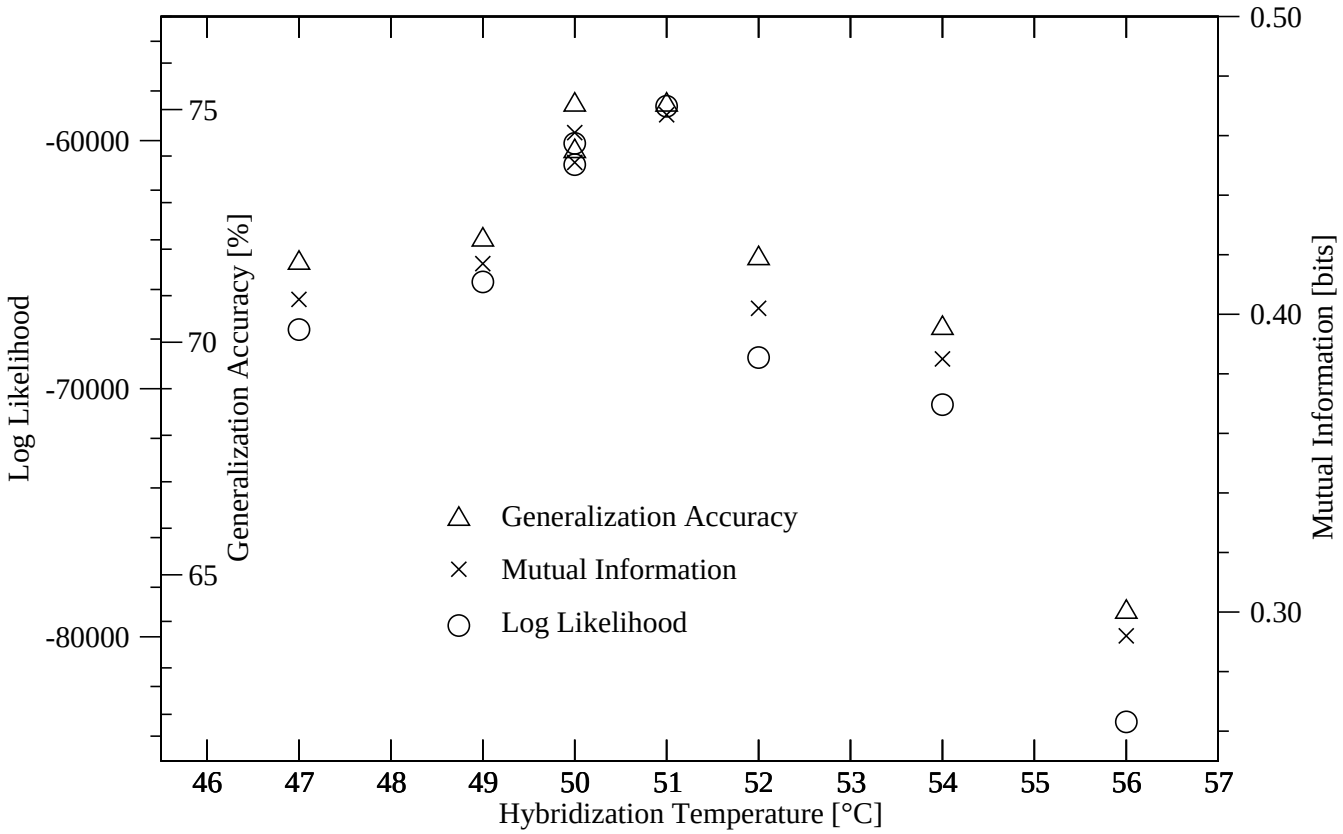

Figure S-2.1: Summary of Performance Measures for Male–Female Separation in response to different hybridization temperatures. This illustration of Table 1 of the main manuscript provides a summary of performance measures for male–female separation.

formation, and the generalization accuracies were constructed by pooling the predictions of all probes. A hybridization temperature of 51°C again gave the best predictive performance, as illustrated in the ROC curves (Fig. S-2.2), agreeing with the optimal temperature derived by analysis of the log likelihoods shown in the main manuscript.

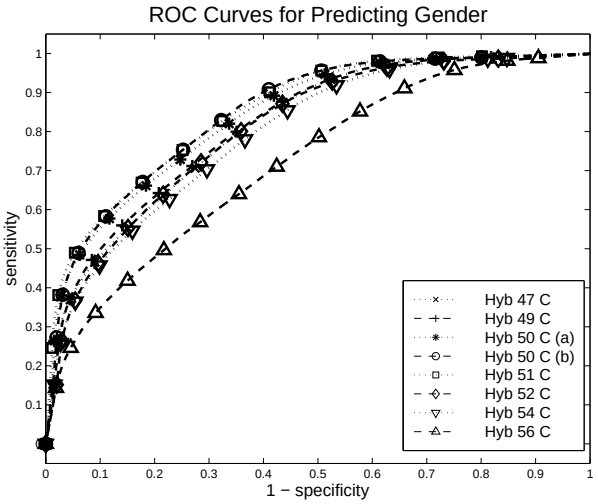

Figure S-2.2: ROC curves for male–female prediction in response to different hybridization temperatures. ROC curves for male–female prediction obtained by six fold cross-testing are shown. All transcripts are used individually to predict whether a channel measurement indicates a male or female sample. We then plot the achieved overall true positive rate (sensitivity) as a function of the accepted overall false positive rate (1 minus specificity). A larger area under the ROC curve corresponds to better performance.

S-2.2.3 Effects of cross-hybridization

We employ the ability of probes to discriminate between two distinct biological samples as an indicator of array performance. To be effective, such a measure needs to reflect that cross-hybridization degrades array performance. Both differentially expressed and non-differentially expressed transcripts can cross-hybridize to probes. Cross-hybridization of differentially expressed targets *adds* unwanted information about sample differences to other, non-target probes. In contrast, cross-hybridization of transcripts that are not differentially expressed with probes that should be detecting differential signals will reduce the overall information about sample differences. If the former effect out-weighs the latter, then the overall information about sample differences will be reduced for conditions with increased cross-hybridization, such

as lower hybridization temperatures. Consistent with this we have shown that the most informative hybridization temperature was not the lowest one tested.

To independently confirm that our results were not affected by cross-hybridization of differentially expressed transcripts we compared the differential expression signal of each probe and that of potential cross-hybridizing non-target probes identified during the probe design process. For every probe with cross-hybridization potential, and for the corresponding non-target probes, we measured the pairwise discrimination performance between target and non-target probes.

The empirical cumulative distribution functions (cdfs) of the log likelihoods (Fig. S-2.3) as well as the receiver operating characteristic (ROC) curves (Fig. S-2.4) both indicate that matching probes are best distinguished from potential cross-hybridizing probes at a physical hybridization temperature of 51°C. This confirmed an absence of bias caused by cross-hybridization of differentially expressed targets, corroborating the robustness of the original calibration analysis of $T_{\mathrm{hyb}}^{(\mathrm{eff})}$ as reported in the manuscript.

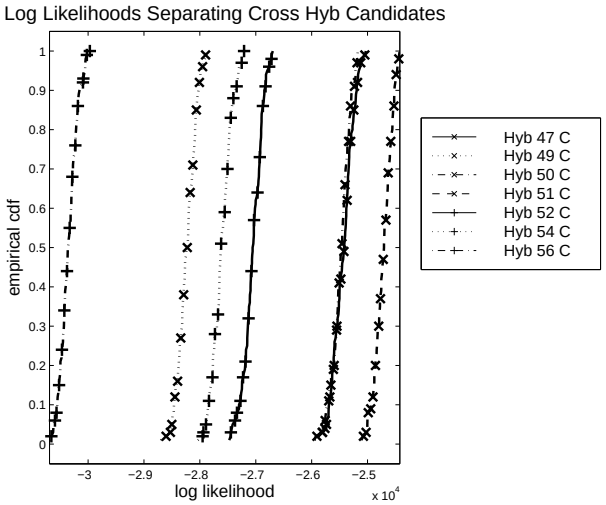

Figure S-2.3: Empirical cdfs of the maximum log likelihood (see main paper for formula) for subsets of 5000 transcripts. The higher the log likelihood, the more evidence the microarray data provide for the sample labels, *i. e.*, the separability of matching probes and potentially cross-hybridizing probes. We again obtained 51°C as optimal hybridization temperature. Also note that the smallest difference between any two experiments is observed for the two independent labelling runs at 50°C indicating a relatively small technical variance.

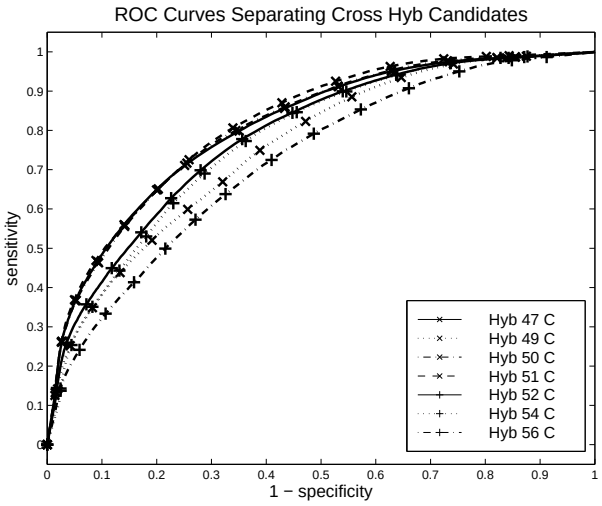

Figure S-2.4: ROC curves for the prediction of a probe being a perfect match or a probe expected to cross-hybridize, obtained by six fold cross testing. All transcripts are used individually for this prediction. We then plot the achieved overall true positive rate (sensitivity) as a function of the accepted overall false positive rate (1 minus specificity). A larger area under the ROC curve corresponds to better performance.

S-2.2.4 Log Ratios, Scatter Plots and Summary Statistics

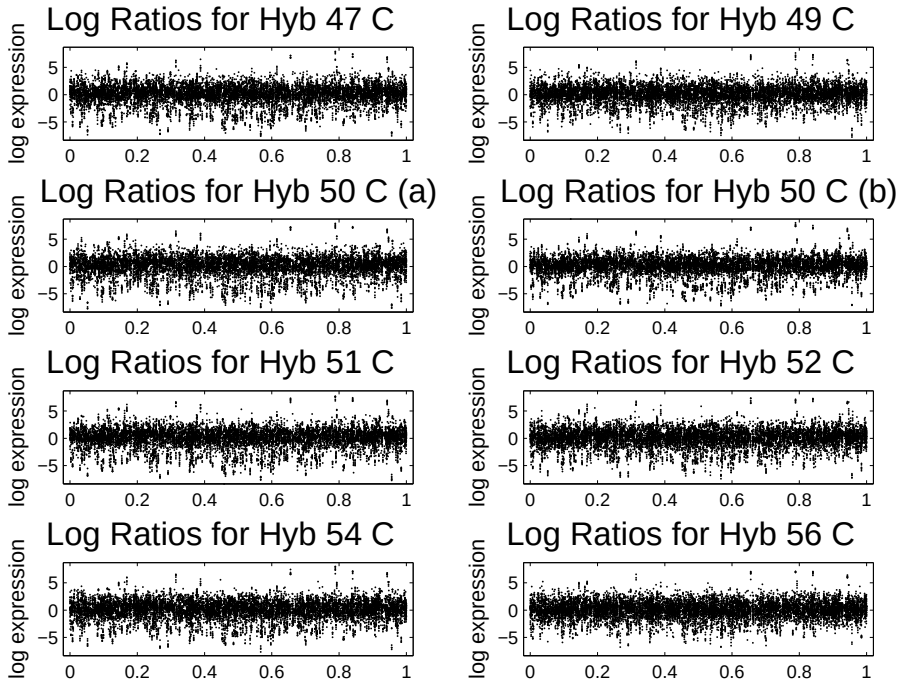

Figure S-2.5: Scatter plots of log ratios for different hybridization conditions. These plots illustrate an identical subset of two-thousand randomly selected genes. For every gene we have set a random x-coordinate that we keep fixed for all hybridization conditions. Despite all the care taken in constructing the plot, it is impossible to assess hybridization conditions coherently, or identify the optimal setting for the hybridization temperature.

To motivate our use of the proposed information theoretic assessment of microarray experiments, we will briefly discuss two apparent alternatives to highlight that measures that might intuitively make sense cannot necessarily be used for quantitative assays. We will first inspect scatter plots of log ratios (*2*) and then examine a summary statistic aiming to capture the separability of male and female *Drosophila* expression measurements.

Figure S-2.5 compares eight plots showing log ratios of genes for different hybridization temperatures. The same subset of genes is illustrated in each panel. By examining a subsample of genes one can still see individual measurements in the plots. The x-coordinates for the genes were drawn randomly once and kept fixed for all panels. The log ratios of individual measurements of the corresponding genes are shown on the y-axis. Although we do not want to anticipate subjective opinions, our conclusion from these plots is that they all look very similar and we cannot identify the best hybridization condition or provide conducive arguments on how one might arrive at quantitative conclusions from simple plots of log ratio data, despite them reflecting differential expression.

A major difficulty in the interpretation of such plots is obviously that the expression measurements are subject to random fluctuations which make it very difficult to assess the values arising from different conditions. Moreover, all genes should be assessed together, for which the plots also do not provide a suitable structure. One might thus argue that a more structured summary statistic might do the trick. For that purpose, we also looked at the empirical cumulative distribution function (cdf) over log ratios, which we plot in Fig. S-2.6. The derivative of a cdf is the corresponding probability density function. The cdf captures thus the same information like a histogram (the integral thereof), however without the difficulty of having to make a trade off between resolution and smoothness. The cdfs are labelled by the hybridization condition, next to which we display the expectation (*i. e.*, the sample mean) of the squared log ratios, $\langle M^2 \rangle$, where expectations are taken over replicate slides and genes.

Looking at the plot we might get a faint idea that the largest hybridization temperatures have a tendency to degrade differential expression. The cdf plots provide however, no means whatsoever for getting the same fine grained and robust resolution that could be obtained with the information theoretic approach introduced in the manuscript. This is in particular evident when considering the expectations over squared log ratios, $\langle M^2 \rangle$ (see figure legend). Firstly, based on this measure, one one would wrongly select

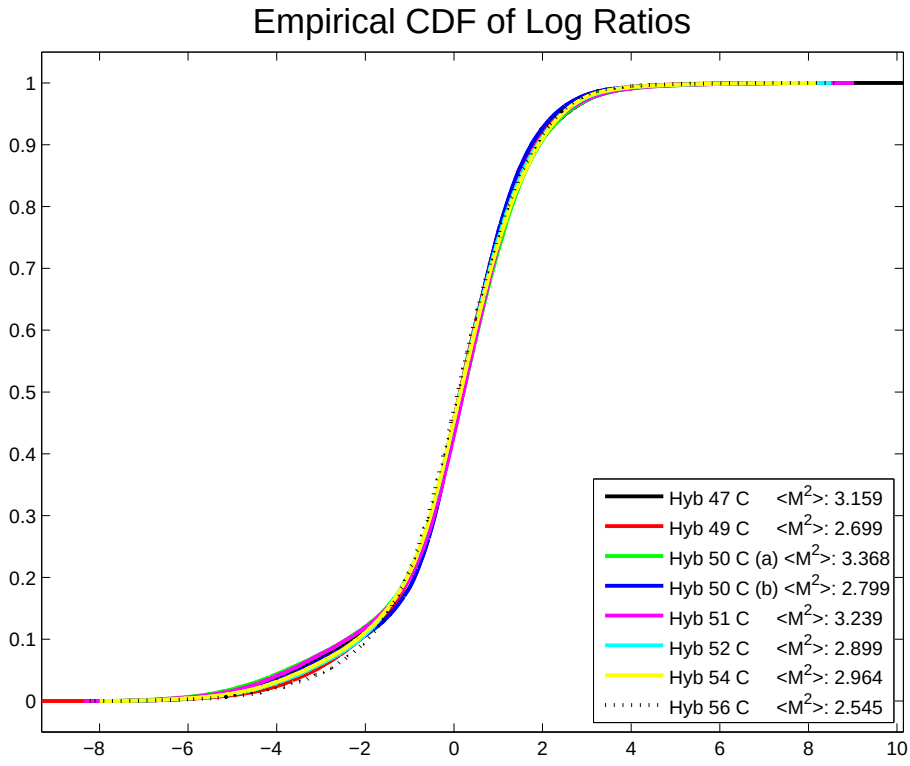

Figure S-2.6: Empirical cdfs of log ratios for different hybridization conditions: While the plot suggests that very large hybridization temperatures tend to degrade log ratios a conclusive ranking is impossible. When inspecting the average squared log ratios, $\langle M^2 \rangle$, we see that simple statistics of log ratios are actually misleading and more volatile than the information theoretic assessment of hybridization conditions proposed in the manuscript.

50°C as the 'best' condition, in disagreement with results for the Shannon channel capacity quantifying the amount of information that can be captured in an experiment, and which is largest for 51°C. Moreover, a ranking by $\langle M^2 \rangle$ is highly volatile, with one of the two identical 50°C runs being ranked first and the other ranked fifth. This example highlights that the information theoretic approach we propose in our work provides a much more robust assessment and that the simpler statistics cannot be relied upon in a selection of the most informative experimental conditions. The information theoretic approach has also conceptual advances, since it provides comple-

mentary measures that allow a diagnosis of potential shortcomings of the estimation procedure. In addition we obtain a universal quantification of the experiment value, telling us how many bits of information about the classification of a biological sample we can per probe on average extract from an experiment.

S-2.3 Temperature dependant gene lists obtained by ANOVA modelling

To augment the GLM based assessment of hybridization protocols introduced in the main paper, we there propose to also consider the length of gene lists obtained by some standard analysis of differential expression. We used the tool FSPMA to this end (*4*), which is a wrapper around a mixed model ANOVA tool originally established by Wernisch et al. (*5*). We find that the lengths of gene lists thresholded with the same cut-off value of $p = 0.05$ correlates well with each of the complementary assessment approaches that we have introduced for the selection of optimal hybridization conditions.

For this analysis we have applied vsn (*3*) for normalizing the array data and calculated p-values for *Drosophila* female RNA being significantly different from male RNA by means of a mixed model ANOVA, properly taking into account nesting effects (technical replication nested within dye.swap). The gene lists obtained for the respective hybridization temperatures can be downloaded by clicking on the respective file name in Table S-2.3. The genes in these tables are ranked with respect to increasing p-values (*i. e.*, decreasing significance). Every row lists an index used for internal purpose, 'dx', a gene symbol, 'g.nams', the p-value which assesses differences in expression for significance, 'p.val', a flag whether the gene is overexpressed in female RNA, 'is.up', and finally the average log fold-change between female and male RNA, 'Cy5.Cy3.mean' (*i. e.*, $\langle \log_2(fem_{RNA}/mal_{RNA}) \rangle$, where $\langle \ldots \rangle$ denotes taking the average).

hyb. temp.	FSPMA rank table
47°C	test.variety.vsn_fly_47_tst.csv
49°C	test.variety.vsn_fly_49_tst.csv
50°C (first replicate)	test.variety.vsn_fly_50a_tst.csv
50°C (second replicate)	test.variety.vsn_fly_50b_tst.csv
51°C	test.variety.vsn_fly_51_tst.csv
52°C	test.variety.vsn_fly_52_tst.csv
54°C	test.variety.vsn_fly_54_tst.csv
56°C	test.variety.vsn_fly_56_tst.csv

Chapter S-3

Methods

This chapter contains the detailed experimental protocols. Additional information is also available on www.flychip.org.uk.

S-3.1 Probe spotting

The probe layout used is available online (FL002 layout).

S-3.2 Slide processing

Slides from commercial sources were processed according to protocols recommended by the manufacturer.

S-3.3 Hybridization Protocol

All hybridizations and washes were performed using an automated hybridization station (Genomic Solutions GeneTAC HybStation).

Sample preparation, labelling

100 μg of total RNA was extracted from male or female *D. melanogaster* following the group's protocol optimized for large scale extraction of RNA from adult flies. RNA quality was verified by electrophoresis and ethidium bromide staining as well as UV spectrometry using an in-house calibrated Nanodrop ND-1000 spectrophotometer. RNA was then labelled by direct incorporation of Cy3-dCTP (Amersham, Cat. No. PA 53021) or Cy5-dCTP (Amersham, Cat. No. PA 55021) in a reverse transcription reaction primed by anchored oligo $(dT)_{23}$ (Sigma, Cat. No. 04387) using Superscript III Reverse Transcriptase (Invitrogen, Cat. No. 18080-044). This was repeated twelve times for each combination of dye and gender, giving a total of $12 \times 2 \times 2 \times 100\,\mu$g of labelled RNA. All male-Cy3 / female-Cy5 samples were then pooled and split into 12 aliquots, sufficient for 24 arrays. Similarly, all female-Cy3 / male-Cy5 samples were pooled and aliquoted. Aliquots were the dried down with a speed vacuum and stored at $-20°$C. Full details of the standard labelling protocol of the group that was used are available online.

Before hybridization, the required number of labelled sample aliquots were resuspended in Ocimum hybridization buffer (Biosolutions, Cat. No. 1180-200000) and sonicated salmon sperm DNA equivalent to 20 μgper array (Invitrogen; Cat. No. 15632-011), pooled, and split into aliquots corresponding to the number of arrays to hybridize.

Hybridization and washes

Hybridization was performed with an automated GeneTAC Hybridization Station (Genomic Solutions). Standard post hybridization washes were performed manually. The group's standard hybridization protocol was employed, with hybridization temperature and duration varied as described in the Methods section of the main manuscript.

S-3.4 Data acquisition and post-processing

Arrays were scanned using a GenePix 4000B dual laser scanner and GenePix Pro 5.1 imaging software (Axon Instruments). Arrays were scanned at 5 μm resolution, simultaneously in both the Cy3 channel (excited by a 532nm laser) and the Cy5 channel (excited by a 635nm laser). Laser power was always set at 100% but photomultiplier tube (PMT) gain was separately adjusted for each channel in order to balance the signal from the two channels and to scan at the highest PMT value for which there were not more than a handful of saturated spots.

S-3.5 Raw data and preprocessing scripts

The data files and preprocessing scripts used are provided, including results from alternative microarray image analysis ('spot finding') tools.

S-3.6 Software

Matlab functions (MatLab® of The MathWorks™) which allow recalculating the evaluations used in the original manuscript are provided at `http://www.sykacek.net/downloads.html#lab_calib`. We provide code as 9.9MB zip archive and as a 9.9MB tgz archive. The zip archive can be unzipped by downloading the file and issuing the command "unzip bmc_code_supp_2010.zip" on a unix type command line. The gzipped tar archive can be expanded issuing the command "tar -xzf bmc_code_supp_2010.tgz". Alternatively the file browsers of most modern operating systems can be used for allocating and expanding these archives as well. After expanding, one finds the new directory ./bmc_code_supp in the current directory. Within this directory the folder ./bmc_code_supp/eval/ contains data files and evaluation scripts which allow repeating all analysis steps concerning the quantitative calibration approach proposed in the paper.

The data files were generated with FSPMA from the raw data (Blue-Fuse quantified images) provided here, using location removal as normalisation method. Different normalisation methods like vsn did not alter the results.

S-3.6.1 Data files for analysis

bfs_fly2_lc_raweffdesc.tsv

contains the effects description which in this case is mainly used for allocating the samples which correspond to different hybridisation temperatures.

bfs_fly2_lc_rawlogG.tsv

contains the expression values of all samples and genes for male Drosophila flies (dye swap resolved).

bfs_fly2_lc_rawlogR.tsv

contains the expression values of all samples and genes for female Drosophila flies (dye swap resolved).

S-3.6.2 MatLab Scripts

There are two MatLab scripts which allow redoing the calculations and evaluations proposed in the paper.

fly_eval_selections.m

allows calculating the proposed quantitative measures which are based on several library functions which are provided in this code supplement. The only additional dependency is introduced by ANOVA based gene rankings which are calculated as well. The latter depend on the MatLab statistics

toolbox. Users without statistics toolbox are advised setting all relevant code fragments under comments. For test purposes it is recommended changing the flag dosubsel=0; to dosubsel=1; for speeding up calculations.

fly_plot_selections.m

generates the tables and plots from the quantitative evaluation measures obtained with fly_eval_selections.m. This function should be called after completion of the evaluations with fly_eval_selections.m.

All code comprising of the above mentioned scripts and the required library functions provided with this code supplement in the directory ../bmc_code_supp/mlablib/ are released under the GPL 2 license. This implies that anyone can modify and use the code under the conditions detailed in this license. An important implication of GPL 2 is that we take no responsibility for wrong conclusions or other damage which might be caused by using the proposed method or software. The library functions provide extensive documentation. This allows adapting the evaluation scripts discussed above to specific user needs.

Application	*Cy3*	*Cy5*
Calibration and validation	550	600

Table S-3.1: PMT gain employed (on a scale 0 . . . 1000)

Bibliography

1. A. M. Dudley, J. Aach, M. A. Steffen, and G. M. Church. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. U. S. A.*, 99(11):7554–7559, 2002. 4

2. T. Han, C. D. Melvin, L. Shi, W. S. Branham, C. L. Moland, P. S. Pine, K. L. Thompson, and J. C. Fuscoe. Improvement in the reproducibility and accuracy of DNA microarray quantification by optimizing hybridization conditions. *BMC Bioinformatics*, 7 Suppl 2:S17, 2006. 11

3. W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1: S96–104, 2002. 14

4. P. Sykacek, R. Furlong, and G. Micklem. A Friendly Statistics Package for Microarray Analysis. *Bioinformatics*, 21:4096–4070, 2005. URL: `http://bioinformatics.oxfordjournals.org/cgi/reprint/bti663?ijkey=UxcmV7ypaHrMDUN&keytype=ref`. 14

5. L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. G. Butcher, and N. G. Stoker. Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, 19(1):53–61, 2003. 14