

MGED Guide to authors, editors and reviewers of microarray gene expression papers

These guidelines are based on the document called Minimum Information About a Microarray Experiment – MIAME [1], developed by the Microarray Gene Expression Data society. For more information see www.mged.org/miame.

Experiment

To biologists, the quintessential information about microarray data concerns the experiment itself. Each of the sections below is required for readers to understand the experiment, to identify the sequences being assayed, and to interpret the resulting data.

Experimental design: This section is quite naturally addressed as part of the results of a publication. Without this type of background information, it is unlikely that anyone could understand the data, which are highly context dependent. Most well-written manuscripts, in addition to containing a general description of the experiment and its goal, will cover the type of experiment, hybridization design, experimental factors, the number of hybridizations performed, the type of reference used for the hybridization, quality control steps taken, and the URL of any supplemental website.

Samples used, extract preparation and labeling: For all samples used in the experiment, information should be given on the origin of the biological sample and its characteristics, manipulation of the samples and the protocols used, protocols for preparing the hybridization and labeling extracts, and any external controls added.

Hybridization procedures and parameters: The protocol and conditions used during hybridization, blocking and washing should be described.

Measurement data and specifications: Every microarray experiment provides three types of measured data: scanned images (raw data); the quantitations based on the images; and the set of quantitations from several arrays upon which the authors base their conclusions. This should be provided on a supplementary website or in an online database in a format that allows further data analysis (e.g., as a spreadsheet or tab-delimited ascii files). While access to images of raw data is not required, authors should make every effort to provide the names (and versions) of scanning hardware and software used, the type of image analysis software used, a description of the measurements produced by the image analysis software and a description of which measurements were used in the analysis. The parameters used with both software and hardware should be stated as well. Because of the complexity of microarray data, how one analyzes and transforms the data can have a profound effect on the results one derives from it. Unlike the case in other high-throughput biological assays, such as DNA sequence, the results are much more dependent on how the data are analyzed. Consequently, both the raw data and the processed results are necessary for an independent reader to validate the conclusions that are drawn from it.

A simple method for presenting this data could be represented as a set of tables. The first could contain the “raw” output of the image analysis software (spot quantitation matrix), the second could contain the “processed” data following normalization and

transformation (gene expression data matrix), and if one is produced, the final table could contain “summary” data that was ultimately used in the analysis, such as the subset of differentially expressed genes identified or gene clusters.

Array Design

Information provided with each array should describe what occupies each feature (spot) on the microarray, as well as the specifications of the array manufacture itself. If an array is not commercially available, the authors should provide sufficient information to allow others to manufacture a similar array. The paper should provide information describing the general array design, including the platform type (whether the array is a spotted glass array, an *in situ* synthesized array, etc.); surface and coating specifications (when known – often commercial suppliers do not provide this data); and the availability of the array (the name or make of commercially available arrays).

For each feature (spot) on the array, its location on the array and the ID of its respective reporter (molecule present on each spot) should be given. For each reporter, its type (e.g., cDNA or oligonucleotide) should be given, along with information that characterizes the reporter molecule unambiguously. This information can be provided in the form of appropriate database reference(s) and sequence (if available), as shown in the example below.

Whenever possible, references should also be provided to the genetic entity to which it maps. For many arrays, the appropriate reference will be to a gene. For example, the cDNA clone IMAGE: 32017 (a reporter) maps to the gene *FNTA* and should be annotated as such. In other cases, reporters may map to ESTs of unknown function or to entities other than genes, such as promoter elements or telomeres, depending on the type of experiment that was done. We refer to the elements to which reporters map as 'composite sequences', because many reporters may be used to assay a single composite sequence. For instance, in the example below, the *ALK* gene is represented on the array by several oligonucleotides. This information can be provided as a spreadsheet or table, either on a supplementary web-site or in one of the abovementioned databases. For more detailed examples see www.mged.org/miame.

The MIAME Checklist

Experiment Design:

- Type of experiment: for example, is it a comparison of normal vs. diseased tissue, a time course, or is it designed to study the effects of a gene knock-out?
- Experimental factors: the parameters or conditions tested, such as time, dose, or genetic variation.
- The number of hybridizations performed in the experiment.
- The type of reference used for the hybridizations, if any.
- Hybridization design: if applicable, a description of the comparisons made in each hybridization, whether to a standard reference sample, or between experimental samples. An accompanying diagram or table may be useful.
- Quality control steps taken: for example, replicates or dye swaps.

- URL of any supplemental websites or database accession numbers

Samples used, extract preparation and labeling:

- The origin of the biological sample (for instance, name of the organism, the provider of the sample) and its characteristics: for example, gender, age, developmental stage, strain, or disease state.
- Manipulation of biological samples and protocols used: for example, growth conditions, treatments, separation techniques.
- Protocol for preparing the hybridization extract: for example, the RNA or DNA extraction and purification protocol.
- Labeling protocol(s).
- External controls (spikes).

Hybridization procedures and parameters:

- The protocol and conditions used during hybridization, blocking and washing.

Measurement data and specifications:

- The quantitations based on the images.
- The set of quantitations from several arrays upon which the authors base their conclusions. While access to images of raw data is not required (although its value is unquestionable), authors should make every effort to provide the following:
 - Type of scanning hardware and software used: this information is appropriate for a materials and methods section.
 - Type of image analysis software used: specifications should be stated in the materials and methods.
 - A description of the measurements produced by the image-analysis software and a description of which measurements were used in the analysis.
 - The complete output of the image analysis *before* data selection and transformation (spot quantitation matrices).
 - Data selection and transformation procedures.
 - Final gene expression data table(s) used by the authors to make their conclusions *after* data selection and transformation (gene expression data matrices).

Array Design:

- General array design, including the platform type (whether the array is a spotted glass array, an *in situ* synthesized array, etc.); surface and coating specifications (when known – often commercial suppliers do not provide this data); and the availability of the array (the name or make of commercially available arrays).
- For each feature (spot) on the array, its location on the array and the ID of its respective reporter (molecule present on each spot) should be given.

- For each reporter, its type (e.g., cDNA or oligonucleotide) should be given, along with information that characterizes the reporter molecule unambiguously, in the form of appropriate database reference(s) and sequence (if available).
- For commercial arrays: a reference to the manufacturer should be provided, including a catalogue number and references to the manufacturer's website if available.
- For non-commercial arrays, the following details should be provided:
 - The source of the reporter molecules: for example, the cDNA or oligo collection used, with references.
 - The method of reporter preparation.
 - The spotting protocols used, including the array substrate, the spotting buffer, and any post-printing processing, including cross-linking.
 - Any additional treatment performed prior to hybridization.

Oligonucleotide array description file example:

| Feature | | | | Reporter | | | | | | Composite sequence | | | |
|----------------------|-------------|-----|-----|---|---------------------|--------------------------|---------------------------|-------------------|-----------------|--------------------|---------------|--|-------------------|
| Coordinates on Array | | | | Reporter ID (user defined) Oligo ID | Biosequence Type | Sequence | DDBJ/ EMBL/ Genbank | Reporter Usage | Control Type | Comp. ID | Designation | Related Gene Symbol, if appropriate | Database Entry |
| Meta Col | Meta Row | Col | Row | | | | | | | | | | |
| 1 | 1 | 1 | 1 | Cy3Cy5 | Oligo | AAAAAAAAAAAA AAAAAA | _ | Control | Positive | C001_01 | Labeled oligo | _ | _ |
| 1 | 1 | 2 | 1 | M00868_01 | Oligo | ACCAGCAGATA CCTCCTTG | D83002 | Experimental | _ | 02_01 | Gene | ALK | LocusID 11682 |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| 4 | 6 | 10 | 8 | M00264_01 | Oligo | ATGTCCGTTGA ATTGG | D83002 | Experimental | _ | C002_01 | Gene | ALK | LocusID 11682 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4 | 6 | 11 | 8 | M02404_01 | Oligo | AGTGGCGAGGA GGAGGAC | L11065 | Experimental | _ | 449_01 | Gene | DPRK1 | LocusID 18387 |
| 4 | 6 | 12 | 8 | M03172_01 | Oligo | CCACCACCAAG ACCTACTCC | U34891 | Experimental | _ | 450_01 | Gene | KLRA9 | LocusID 16640 |

cDNA array description file example:

| Feature | | | | Reporter | | | | | | Composite sequence | | | |
|----------------------|-------------|-----|-----|---|---------------------|------------------|---------------------------|-------------------|-----------------|--------------------|-------------|---------------------------|-------------------|
| Coordinates on Array | | | | Reporter ID (user defined) HGMP Ref | Biosequence Type | Clone ID | DDBJ/ EMBL/ Genbank | Reporter Usage | Control Type | Comp. ID | Designation | Related Gene Symbol | Database Entry |
| Meta Col | Meta Row | Col | Row | | | | | | | | | | |
| 1 | 1 | 1 | 1 | 370503 | cDNA clone | IMAGE 32017 | R17905 | Experimental | _ | C1 | Gene | FNTA | LocusID2339 |
| 1 | 1 | 2 | 1 | 370504 | cDNA clone | IMAGE 2962831 | BC005866 | Experimental | _ | C2 | Gene | MLH1 | LocusID 4292 |
| 1 | 1 | 3 | 1 | 370505 | Genomic clone | Cosmid 9H11 | L40416 | Control | Positive | _ | _ | _ | _ |
| : | : | : | : | : | : | : | : | : | : | : | : | : | : |
| 4 | 8 | 24 | 12 | 380696 | cDNA clone | IMAGE 5214483 | BC028215 | Experimental | _ | C285 | Gene | PTEN | LocusID 5728 |

Summary

Our hope is that these guidelines are simple enough that they can be applied by all generators of microarray data, yet are comprehensive enough that they will allow both journals and their readers to put published data to its best use. We believe that the template provided here can become the basis for a standardized presentation of microarray results. However, the only practical standards are those that are widely endorsed and accepted. Consequently, we look for input from both the scientific journals and the community as to the feasibility of applying these standards. To facilitate the discussion we have created an e-mail discussion group (microarray annotations) that can be joined via www.mged.org.

References

[1] A. Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, C A Ball, H C Causton, T Gaasterland, P Glenisson, F C P Holstege, I F Kim, V Markowitz, J C Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo & M Vingron. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data, *Nature Genetics*, vol 29 (December 2001), pp 365 - 371.