

.mev File Format Description

Production Revision 4.0 (July 16, 2004)

Description of MultiExperimentViewer files (.mev)

A MultiExperimentViewer or .mev file is a tab-delimited text file that contains coordinate and expression data for a single microarray experiment. A single header row is required to precede the expression data in order to identify the columns below. With the exception of optional comment lines, each remaining row of the file stores data for a particular spot/feature on the array.

MeV and other TM4 software tools will consider comment lines non-computational. A comment line must start with the pound symbol '#', and can be included anywhere in the file. If the pound symbol is the first character on a line, the entire line (up to the newline character '\n') will be ignored by the software tool.

The mev files created at TIGR will typically contain at least one comment at the top of the file with the following information. This information is optional to non-TIGR users. The format and fields contained within these comments are subject to change. See **Appendix 2** for details.

| | |
|------------------|--|
| version | Version number based on revisions of expression data |
| format_version | The version of the .mev file format document |
| date | Date of file creation or update |
| analyst | Owner or the person responsible for creating the file |
| analysis_id | <i>id</i> from the <i>analysis</i> table that corresponds to this set of expression values |
| slide_type | <i>slide_type</i> from the <i>slide_type</i> table that this array is based on |
| input_row_count | Number of rows of expression (eg. non-header) data in input files |
| output_row_count | Number of rows of expression (eg. non-header) data in this file |
| created_by | Software tool used to create the file |
| description | Common name or other details about the experiment |

An example of the leading comments:

```
# version: V1.0
# format_version: V4.0
# date: 10/06/2004
# analyst: aisaeed
# analysis_id: 10579
# slide_type: IASCAG1
# input_row_count: 32448
# output_row_count: 32448
# created_by: TIGR Spotfinder 2.2.3
# TIFF files processed: gpc30025a_532_nm.tif, gpc30025a_635_nm.tif
# description: Tumor type comparison
# This is the 4th experiment in a series of 20 to identify tissue-specific genes.
```

The header row consists of the field names for each subsequent row in this file (with the exception of comment lines). A minimum of seven columns must be present, and these must use a set of specifically named headers. Any number of additional columns may be included. The seven required column headers are:

| | |
|-----|---------------------------------|
| UID | Unique identifier for this spot |
| IA | Intensity value in channel A |
| IB | Intensity value in channel B |
| R | Row (slide row) |
| C | Column (slide column) |
| MR | Meta-row (block row) |
| MC | Meta-column (block column) |

As of version 4.0 of this file format the IA and IB columns can be substituted with MedA and MedB. The new requirement is that at least one integrated intensity (IA, IB, etc.) **or** one median (MedA, MedB, etc.) value be reported for each channel in the microarray. For example, a two channel microarray .mev file would require either IA and IB **or** MedA and MedB.

| | |
|------|-------------------------------|
| MedA | Median intensity in channel A |
| MedB | Median intensity in channel B |

The mev files created at TIGR may use one of the following formats for the header row, depending on the origin of the mev file. The non-required columns (i.e. anything after the 7th column) may be rearranged and their names are subject to change at this time.

1) Database created mev file:

```
UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA \t SAB \t SFA \t SFB \t
QCS \t QCA \t QCB \t BkgA \t BkgB
```

| | |
|-------|--|
| UID | Unique identifier for this spot |
| IA | Intensity value in channel A |
| IB | Intensity value in channel B |
| R | Row (slide row) |
| C | Column (slide column) |
| MR | Meta-row (block row) |
| MC | Meta-column (block column) |
| SR | Sub-row |
| SC | Sub-column |
| FlagA | <i>TIGR Spotfinder</i> flag value in channel A |
| FlagB | <i>TIGR Spotfinder</i> flag value in channel B |
| SAA | Actual spot area (in pixels) in channel A |
| SAB | Actual spot area (in pixels) in channel B |
| SFA | Saturation factor in channel A |
| SFB | Saturation factor in channel B |

| | |
|------|------------------------------------|
| QC | Cumulative quality control score |
| QCA | Quality control score in channel A |
| QCB | Quality control score in channel B |
| BkgA | Background value in channel A |
| BkgB | Background value in channel B |

2) Spotfinder created mev file:

```
UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA \t SAB \t SFA \t SFB \t
QCS \t QCA \t QCB \t BkgA \t BkgB \t SDA \t SDB \t SDBkgA \t SDBkgB \t MedA \t MedB \t
AID
```

| | |
|--------|---|
| UID | Unique identifier for this spot |
| IA | Intensity value in channel A |
| IB | Intensity value in channel B |
| R | Row (slide row) |
| C | Column (slide column) |
| MR | Meta-row (block row) |
| MC | Meta-column (block column) |
| SR | Sub-row |
| SC | Sub-column |
| FlagA | <i>TIGR Spotfinder</i> flag value in channel A |
| FlagB | <i>TIGR Spotfinder</i> flag value in channel B |
| SAA | Actual spot area (in pixels) in channel A |
| SAB | Actual spot area (in pixels) in channel B |
| SFA | Saturation factor in channel A |
| SFB | Saturation factor in channel B |
| QC | Cumulative quality control score |
| QCA | Quality control score in channel A |
| QCB | Quality control score in channel B |
| BkgA | Background value in channel A |
| BkgB | Background value in channel B |
| SDA | Standard deviation for spot pixels in channel A |
| SDB | Standard deviation for spot pixels in channel B |
| SDBkgA | Standard deviation of the background value in channel A |
| SDBkgB | Standard deviation of the background value in channel B |
| MedA | Median intensity value in channel A |
| MedB | Median intensity value in channel B |
| AID | Alternative ID |

The first seven fields (UID, IA, IB, R, C, MR and MC) are required as specified above.

This flexible format allows users to track slide-specific data of interest, such as background, spot size and alternate intensities without requiring them of all users or adopting a limited ‘vocabulary’ of field names. This header row serves to identify the required and additional data

columns. UID must be the left-most column in the mev file. Other columns do not need to be present in a fixed order.

For mev files generated at TIGR, the UIDs may be of the form: *database_name:spot_id* (eg. cage:20238). For any given microarray database, the *id* field in the *spot* table will be unique. The combination of database and *spot_id* will therefore uniquely identify any spot on any array created at TIGR. It is important to note that this is not enough information to distinguish between spots in the same location on two slides of the same *slide_type*, as this would typically require an *analysis_id*. Since annotation data is based on *slide_type*, it is not necessary to make this distinction, as all slides of a given type will use the same annotation file.

The AID column will usually contain an incremental sequence of numbers starting at 1. These can be used to return the file to the original sorted order and can function as a unique row identifier if necessary.

Applications that generate files of expression data (commonly in *tav* format) by retrieving records from the database access the *spot* table. *TIGR Spotfinder*, *Midas* and *Madam* are all capable of generating UIDs of the form described above in addition to the typical coordinate and intensity data.

mev files are required to end with the extension '.mev'. At this time there are no further naming conventions for mev files.

Description of Annotation Files

An annotation file is a tab-delimited text file containing annotation data for a specific *slide_type*. mev files can be associated with an annotation file only if both types of files are based on the same *slide_type*. The keys to this association are the unique ids in both files. Rows of mev and annotation files can be associated with each other if the unique ids are identical. A single header row is required to precede the annotation data in order to identify the columns below. Each remaining row of the file stores annotation data for a particular spot/feature on the array.

Annotation files may contain any number of non-computational comment lines. These lines, starting with '#', will be treated identically to comment lines in mev files, and should precede the header row.

Annotation files created at TIGR will use UIDs that match the format used in the mev files, most likely *database_name:spot_id*. The structure of each annotation file is detailed below.

The header row consists of headers that identify each column of data. Each subsequent row of the file stores data for a particular spot/feature on the array. The annotation files created at TIGR will typically contain at least one comment at the top of the file with the following information:

| | |
|----------------|---|
| version | Version number based on revisions of annotation data |
| format_version | The version of the .mev file format document |
| date | Date of file creation or update |
| analyst | Owner or the person responsible for creating the file |

| | |
|------------------|--|
| created_by | Software tool used to create the document |
| gi_version | Version of the Gene Indices (or db?) that produced this annotation data |
| slide_type | <i>type</i> from the <i>slide_type</i> table that this array is based on |
| output_row_count | Number of rows of annotation (eg. non-header) data |
| description | Common name or other details about the experiment |

An example of the leading comments:

```
# version: V3.0
# format_version: V4.0
# date: 04/20/2004
# analyst: jwhite
# created_by: Database script
# gi_version: 3.0
# slide_type: IASCAG1
# output_row_count: 32448
# description: Standard annotation file
```

The header row consists of the field names for each subsequent row in this file. Only the UID field is required. It must be the first field present and it must be named 'UID'. Any number of additional fields may be included. Annotation files created at TIGR will always contain the following columns:

```
UID  unique identifier for this line of annotation
R    row (slide row)
C    column (slide column)
```

The remaining fields may vary, and a standard set has yet to be determined. Such a list will be published on a future date. R and C have been included to allow for manual alignment of the mev and corresponding annotation files in the event that the mev files were not generated in a traditional manner (ie. using *Madam*, etc.).

Some varieties of annotation files follow. The format may vary depending on the purpose of the file:

```
UID \t R \t C \t FeatN \t GBNum \t TCNum \t ComN \t ...
UID \t R \t C \t GeneN \t Rxn \t PathwayN \t ...
UID \t R \t C \t FeatN \t End5 \t End3 \t ChrNum \t ...
```

Of course, it would be possible to combine the fields of these files, or add fields that have not been mentioned here. The goal is to keep the annotation flexible and the processing seamless.

There are not any naming conventions for annotation files at this time. If such a standard is introduced in the future, it will be detailed here.

Example MeV File

Based on this format description, the first few rows of a mev file created at TIGR might look like:

```
# version: V1.0
# format_version: V4.0
# date: 10/06/2004
# analyst: aisaeed
# analysis_id: 10579
# slide_type: IASCAG1
# input_row_count: 32448
# output_row_count: 32448
# created_by: TIGR Spotfinder 2.2.3
# TIFF files processed: gpc30025a_532_nm.tif, gpc30025a_635_nm.tif
# description: Tumor type comparison
# This is the 4th experiment in a series of 20 to identify tissue-specific genes.
UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA
cage:1043 \t 20934 \t 390823 \t 1 \t 1 \t 1 \t 1 \t 1 \t 1 \t 1 \t C \t C \t 215 \t 225
cage:1044 \t 298734 \t 90823 \t 1 \t 2 \t 1 \t 1 \t 1 \t 1 \t 2 \t C \t C \t 198 \t 176
cage:1045 \t 789435 \t 713952 \t 1 \t 3 \t 1 \t 1 \t 1 \t 1 \t 3 \t C \t C \t 255 \t 244
```

To the software, the same file would appear to be:

```
UID \t IA \t IB \t R \t C \t MR \t MC \t SR \t SC \t FlagA \t FlagB \t SAA
cage:1043 \t 20934 \t 390823 \t 1 \t 1 \t 1 \t 1 \t 1 \t 1 \t 1 \t C \t C \t 215 \t 225
cage:1044 \t 298734 \t 90823 \t 1 \t 2 \t 1 \t 1 \t 1 \t 1 \t 2 \t C \t C \t 198 \t 176
cage:1045 \t 789435 \t 713952 \t 1 \t 3 \t 1 \t 1 \t 1 \t 1 \t 3 \t C \t C \t 255 \t 244
```

In the event an application would need to gain access to the comment lines, the file parser has methods to allow it.

Example Annotation File

Based on this format description, the first few rows of an annotation file created at TIGR might look like:

```
# version: V3.0
# format_version: V4.0
# date: 04/20/2004
# analyst: jwhite
# created_by: Database script
# gi_version: 3.0
# slide_type: IASCAG1
# output_row_count: 32448
# description: Standard annotation file
UID \t R \t C \t CloneN \t GBNuM \t TCNum \t Guess
cage:1043 \t 1 \t 1 \t A.t.RCA \t M86720 \t null \t null
cage:1044 \t 1 \t 2 \t Image:511428 \t AA126115 \t THC1324489 \t TC: FXYD domain-
containing ion transport regulator 3 precursor
cage:1045 \t 1 \t 3 \t Image:897987 \t AA598884 \t THC1286273 \t TC: NADH-ubiquinone
oxidoreductase 39kDa subunit {Homo sapiens}
```

To the software, the same file would appear to be:

```
UID \t R \t C \t CloneN \t GBNuM \t TCNum \t Guess
cage:1043 \t 1 \t 1 \t A.t.RCA \t M86720 \t null \t null
cage:1044 \t 1 \t 2 \t Image:511428 \t AA126115 \t THC1324489 \t TC: FXYD domain-
containing ion transport regulator 3 precursor
cage:1045 \t 1 \t 3 \t Image:897987 \t AA598884 \t THC1286273 \t TC: NADH-ubiquinone
oxidoreductase 39kDa subunit {Homo sapiens}
```

In the event an application would need to gain access to the comment lines, the file parser has methods to allow it.

Appendix 1: Column Header names and descriptions for .mev files

Note: Deprecated header names are indicated in (*italics*) next to the preferred name. These will be considered deprecated in all new versions of TM4 software.

| Header | Description |
|--------------------------|--|
| AID | Alternative ID (used when UID is substituted) |
| Bkg | Background for one/both channels, if only one is available |
| BkgA (<i>BG1, BGA</i>) | Spot background integral for channel A |
| BkgB (<i>BG2, BGB</i>) | Spot background integral for channel B |
| C | Column of slide |
| ChrNum | Chromosome number |
| CloneID | id from table clone |
| CloneN | Clone name |
| ComN | guess or com_name, i.e. putative function information |
| End3 | 3' end position of a sequence |
| End5 | 5' end position of a sequence |
| ExptN | Experiment Name |
| FeatN | Feature name, i.e. feat_name |
| Flag | Flag for one/both channels, if only one is available |
| FlagA (<i>Flag1</i>) | Flag of channel A |
| FlagB (<i>Flag2</i>) | Flag of channel B |
| GBNum | Genbank number |
| GeneID | id from table gene |
| GeneN | Gene name or symbol |
| GO | GO Term |
| IA (<i>I1</i>) | Spot integral in channel A (Intensity A) |
| IB (<i>I2</i>) | Spot integral in channel B (Intensity B) |
| Locus | Locus |
| MC | Meta column of slide |
| MeanAB | Ratio of spot means (channel A / channel B) |
| MedBkgA | Background median value in channel A |
| MedBkgB | Background median value in channel B |
| MedAB | Ratio of spot medians (channel A / channel B) |
| MedA | Median intensity in channel A |
| MedB | Median intensity in channel B |
| MNA | Mean intensity in channel A |
| MNB | Mean intensity in channel B |
| ModeAB | Ratio of spot modes (channel A / channel B) |
| MR | Meta row of slide |
| NIA | Normalized intensity from channel A |
| NIB | Normalized intensity from channel B |
| OligoID | id from table oligo |
| PathwayN | Pathway name |
| PValue | P Value (for ttest results) |
| QC (<i>QCscore</i>) | Combined quality control score from channels A and B |

| | |
|--------------------|--|
| QCA (<i>QC1</i>) | Quality control score in channel A |
| QCB (<i>QC2</i>) | Quality control score in channel B |
| R | Row of slide |
| RatioAB | Ratio of intensities (channel A / channel B) |
| Rxn | Reaction name or description |
| SA | Spot Area for one/both channels, if only one is available |
| SAA | Spot Area for channel A |
| SAB | Spot Area for channel B |
| SF | Spot saturation factor for both channels i.e. (number of used non-saturated pixels)/(total spot pixels) |
| SFA | Spot saturation factor for channel A only (for GenePix data compatibility) |
| SFB | Spot saturation factor for channel B only (for GenePix data compatibility) |
| SC | Columns of a grid (block) |
| SD | Standard deviation |
| SDBkgA | Standard deviation for background pixels in channel A |
| SDBkgB | Standard deviation for background pixels in channel B |
| SDA | Standard deviation for spot pixels in channel A |
| SDB | Standard deviation for spot pixels in channel B |
| SeqN | Sequence name, i.e. seq_name |
| SID | Spot ID |
| StudyN | Study Name |
| SR | Row of a grid (block) |
| TCNum | TC number (from Gene Indexes) |
| TTestDF | Degrees of Freedom (for ttest results) |
| TTestMean | Mean (for ttest results) |
| TTestSD | Standard Deviation (for ttest results) |
| TValue | t Value (for ttest results) |
| UGNum | Unigene number (from NCBI unigene) |
| UID | Unique Identification |
| ZScore | Z-score |

Appendix 2: Keys for Header Comments for .mev files

| | |
|------------------|---|
| analysis_id | ID from table analysis (for expression data) |
| analyst | Person who created the document |
| created_by | Software tool used to create the document |
| date | Date document was created |
| description | Text description of the algorithm, program or manipulation of the input data. |
| format_version | The version of the .mev file format document |
| input_row_count | Row count of input data files (if applicable) |
| output_row_count | Row count of THIS data file (if available) |
| slide_type | For annotation files |
| source_files | Input data files, e.g. *.tif, *.tav *.mev |
| version | Document version number |