mapDIA: Model-based Analysis of Quantitative Mass Spectrometry Data in Data Independent Acquisition Model

Guoshou Teo and Hyungwon Choi

November 23, 2017

1 Installation

1.1 Linux/macOS

Type "make -j" to compile the software. The source code requires standard g++ for compilation. Due to limitations of GNU Make, the path of mapDIA is not allowed to contain any whitespace characters.

1.2 Windows

Executables for 64-bit Windows are included. Windows users will need to convert text files in the "examples" folder using unix2dos `https://waterlan.home.xs4all.nl/dos2unix/dos2unix-7.4.0-win64.zip`.

2 Input Parameters

mapDIA expects two mandatory input files: the **raw data containing fragment-level (or peptide-level or protein-level) peak area data**, and the **input parameter file** (example files can be found in the "examples" folder). The raw data file has to be formatted according to the instruction provided in the section 3, and the input parameters must be specified according to the user's preference for various aspects of the data processing and modeling.

It is important to specify an appropriate experimental design in mapDIA analysis: the options are (i) independent sample comparison (IS design) and (ii) within-replicate comparison for biological replicates (REP design). In the former, groups of independent biological samples are compared for their difference in mean abundance. In the latter, by contrast, different conditions are compared within each biological sample (or replicate) over at least two biological replicates. To make an analogy to the traditional statistical hypothesis testing,

the model for the IS design corresponds to the two sample independent t-test, whereas the model for the REP design corresponds to the paired t-test.

The mapDIA software requires that the user provide the following options for data processing and modeling:

- **FILE**: The name of the tab-delimited file containing the fragment-level / peptide-level / protein-level peak area data. You can also provide the full path to the file in case the file is located in distant folders. Missing observation can be represented by "0", "NA" or an empty string. Missing retention times can be represented by "NA" or an empty string. Rows with missing retention times will be removed from the analysis if NORMALIZATION=RT. Duplicated row names will have "_duplicate" appended to them in the output files.

- **INCLUSION_LIST**: The list of proteins IDs for which the corresponding proteins are exempt from all filtering steps. If the impute option is not set by the user, these proteins will have missing values set to $0.5\times$(row minimum). This is to ensure sufficient data for further analysis.

- **LEVEL**: Set LEVEL=2 for peptide-level data. Set LEVEL=1 for protein-level data. Set LEVEL=3 or exclude this line for fragment-level data.

- **FUDGE**: A probability for the sample quantiles of the set of non-zero values

$$\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y^2-(\frac{1}{n_A+n_B+1/V})(\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y)^2$$

from all peptides to be set as the fudge factor,y_{fudge}. (e.g. "FUDGE=0.10" will set the fudge factor to the 10th percentile of $\{\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y^2-(\frac{1}{n_A+n_B+1/V})(\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y)^2 : \sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y^2-(\frac{1}{n_A+n_B+1/V})(\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y)^2\neq 0, q\in$ all peptides in the data$\}$) Valid values should be between 0 and 1. Exclude line if fudge factor is not needed.

The full closed form expression of conditional marginal likelihoods with fudge factor incorporated. For $z_p=0$,

$$\begin{aligned}
&\pi(\mathbf{y}_q^A,\mathbf{y}_q^B|z_p=0)\\
=&\int_0^\infty\int_{-\infty}^\infty \varphi(\mathbf{y}_q^A,\mathbf{y}_q^B|\mu_q,\sigma_q^2)\varphi(\mu_q|0,\sigma_q^2V)\mathcal{IG}(\sigma_q^2|a,b)\,d\mu_q d\sigma_q^2\\
\approx&\frac{1}{\sqrt{(n_A+n_B)V+1}}\frac{\Gamma(a+(n_A+n_B)/2)}{\Gamma(a)}\frac{1}{(2\pi)^{(n_A+n_B)/2}}\\
&\times\frac{b^a}{\left[b+\frac{1}{2}\left(\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y^2-(\frac{1}{n_A+n_B+1/V})(\sum_{y\in\mathbf{y}_q^A,\mathbf{y}_q^B} y)^2\underbrace{+y_{\text{fudge}}}\right)\right]^{a+(n_A+n_B)/2}}
\end{aligned}$$

where $\mathcal{IG}(\cdot)$ denotes the inverse gamma density function. Likewise, the closed form expression for the case $z_p = 1$,

$$
\begin{aligned}
&\pi(\mathbf{y}_q^A, \mathbf{y}_q^B | z_p = 1) \\
&= \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \varphi(\mathbf{y}_q^A | \mu_{qA}, \sigma_q^2) \varphi(\mathbf{y}_q^B | \mu_{qB}, \sigma_q^2) \\
&\quad \times \varphi(\mu_{qA} | 0, \sigma_q^2 V) \varphi(\mu_{qB} | 0, \sigma_q^2 V) \mathcal{IG}(\sigma_q^2 | a, b)\, d\mu_{qA} d\mu_{qB} d\sigma_q^2 \\
&\approx \frac{1}{\sqrt{n_A V + 1}} \frac{1}{\sqrt{n_B V + 1}} \frac{\Gamma(a + (n_A + n_B)/2)}{\Gamma(a)} \frac{1}{(2\pi)^{(n_A + n_B)/2}} \\
&\quad \times \frac{b^a}{\left[b + \frac{1}{2} \left(SS_{qA} - (\frac{1}{n_A + 1/V})(S_{qA})^2 + SS_{qB} - (\frac{1}{n_B + 1/V})(S_{qB})^2 \underbrace{+ y_{\text{fudge}}}_{} \right) \right]^{a + (n_A + n_B)/2}}
\end{aligned}
$$

where $SS_{qA} = \sum_{y \in \mathbf{y}_q^A} y^2$, $S_{qA} = \sum_{y \in \mathbf{y}_q^A} y$, $SS_{qB} = \sum_{y \in \mathbf{y}_q^B} y^2$, $S_{qB} = \sum_{y \in \mathbf{y}_q^B} y$, and $n_g = \sum_{y \in \mathbf{y}_q^g} I\{\text{y observed}\}$ is the number of observed intensities in peptide q group g.

- **LOG2_TRANSFORMATION**: Set LOG2_TRANSFORMATION=false if the input data need not be $\log_2$-transformed. LOG2_TRANSFORMATION=true by default.

- **REMOVE_SHARED_PEPTIDE**: Set REMOVE_SHARED_PEPTIDE=true if peptides belonging to more than one protein are to be removed. REMOVE_SHARED_PEPTIDE=false by default. This option is valid only for peptide-level and fragment-level data.

- **IMPUTE**: Set the imputation type and scale. If "IMPUTE=row n", missing values will be imputed as $n\times$ (smallest value for the row), where $n > 0$. If "IMPUTE=group n", missing values will be imputed as $n\times$ (smallest value for the group in the row), where $n > 0$. In the case where there are no observations for a group in a row, missing values will be imputed as $n\times$ (smallest value for the column). Exclude this line if imputation is not required.

- **EXPERIMENTAL_DESIGN**: The type of statistical model. Options are "ReplicateDesign" for the REP design and "IndependentDesign" for the IS design.

- **NORMALIZATION**: Exclude this line if no normalisation is needed. Options are "TIS" for division by the total ion chromatogram, and "RT δ" (δ a positive real number) for retention time-based normalization method described in our paper (e.g. "RT 10" for the Gaussian kernel weights with standard deviation of 10). "RT" option requires an extra column of retention time data (in minutes) appended to FILE. Rows with missing retention time data will be removed. Optionally, append an non-negative integer after a whitespace to indicate the number of decimal places to round off the retention times if needed (this is to speed up the RT normalization)(e.g. "RT 10 3" to round off the retention times to 3 decimal places).

- **Fragment filtering options (for fragment-level data)**: The filtering steps for selecting reliable fragments in mapDIA. In the software, the raw data are first $\log_2$-transformed, and centered by the overall median value in each fragment. The median value is calculated by the median across all samples in the IS design, whereas it is calculated by the median within each biological sample/replicate in the REP design.
  - **SDF**: Following the centering step, fragment peak area value will be removed in a sample if the $\log_2$ peak area of that fragment substantially deviates from the median peak area value across all fragments in the mother protein in the same sample/condition by a specified margin. For each protein, this margin is set as the standard deviation across the fragments times the SDF parameter (**S**tandard **D**eviation **F**actor). For example, if SDF = 2, then the data points lying out of median ± 2 standard deviations in each sample will be discarded (but not other values of the same fragments in other samples/conditions). Set SDF=inf to switch off this filter.
  - **PSEUDOCV**: The pseudo coefficient of variation is defined as the coefficient of variation pretending as if all samples are biological replicates – representing the overall variation of each fragment across the samples. The minimum average correlation filter (**MIN_CORREL**) will only be applied to fragments/peptides in each protein with average pseudoCV above the threshold set by the user. Default is 0, which implies **MIN_CORREL** will be applied to entire data.
  - **MIN_CORREL**: The median intra-protein correlation cutoff. Values should be between -1 and 1.
  - **MIN_OBS**: The list of minimal number of non-missing value for each group in the fragment, if using REP design, one value is sufficient. Values should be between 1 and the available number of samples in the group.
  - **MIN_FRAG_PER_PEP**: the minimum number of fragments per peptide satisfying "MIN_OBS" criteria. Values should be ≥ 1.
  - **MAX_FRAG_PER_PEP**: the maximum number of fragments per peptide satisfying "MIN_OBS" criteria. Set to "inf" to switch off this filter. Values should be $\geq$MIN_FRAG_PER_PEP.
  - **MIN_PEP_PER_PROT**: the minimum number of peptides (with at least MIN_FRAG_PER_PEP fragments) per protein. Values should be ≥ 1.
  - **MAX_PEP_PER_PROT**: The maximum number of peptides to be reported in "protein_level.txt". Set to "inf" to include all peptides used in the analysis. Values should be $\geq$MIN_PEP_PER_PROT.

- **Peptide filtering options (for peptide-level data)**: The filtering steps for selecting reliable peptides in mapDIA. In the software, the raw data are first $\log_2$-transformed, and centered by the overall median value in each peptide. The median value is calculated by the median across all samples in the IS design, whereas it is calculated by the median within each biological sample/replicate in the REP design.

- **SDF**: Following the centering step, peptide peak area value will be removed in a sample if the $\log_2$ peak area of that peptide substantially deviates from the median peak area value across all peptides in the mother protein in the same sample/condition by a specified margin. For each protein, this margin is set as the standard deviation across the peptide times the SDF parameter (**S**tandard **D**eviation **F**actor). For example, if SDF = 2, then the data points lying out of median ± 2 standard deviations in each sample will be discarded (but not other values of the same peptide in other samples/conditions). Set SDF=inf to switch off this filter.
- **PSEUDOCV**: The pseudo coefficient of variation is defined as the coefficient of variation pretending as if all samples are biological replicates – representing the overall variation of each fragment across the samples. The minimum average correlation filter (**MIN_CORREL**) will only be applied to fragments/peptides in each protein with average pseudoCV above the threshold set by the user. Default is 0, which implies **MIN_CORREL** will be applied to entire data.
- **MIN_CORREL**: The median intra-protein correlation cutoff. Values should be between -1 and 1.
- **MIN_OBS**: The list of minimal number of non-missing value for each group in the peptide, if using REP design, one value is sufficient. Values should be between 1 and the available number of samples in the group.
- **MIN_PEP_PER_PROT**: the minimum number of peptides per protein satisfying "MIN_OBS" criteria. Values should be ≥ 1.
- **MAX_PEP_PER_PROT**: the maximum number of peptides per protein satisfying "MIN_OBS" criteria. The maximum number of peptides to be reported in "protein_level.txt". Set to "inf" to switch off this filter. Values should be $\geq$MIN_PEP_PER_PROT.

- **Protein filtering options (for protein-level data)**: The filtering steps for selecting reliable proteins in mapDIA. In the software, the raw data are first $\log_2$-transformed, and centered by the overall median value in each protein. The median value is calculated by the median across all samples in the IS design, whereas it is calculated by the median within each biological sample/replicate in the REP design.
  - **MIN_OBS**: The list of minimal number of non-missing value for each group in the protein, if using REP design, one value is sufficient. Values should be between 1 and the available number of samples in the group.

- **Sample information**:
  - **LABEL**: a list of labels for conditions (no white space within a label)
  - **SIZE**: the corresponding number of samples from each group (or condition), if using REP design, one value is sufficient

- **MIN_DE, MAX_DE**: The minimum and maximum proportion of proteins classified DEPs (i.e. with score > 0.5). Values should satisfy $0 <$ MIN_DE $<$ MAX_DE < 1.

- **CONTRAST**: a binary (0/1) square matrix indicating the comparisons required for the analysis. Note that mapDIA's architecture is designed for multi-group comparisons (more than two groups), and the user is responsible for specifying which pairs of groups are to be compared. For example, suppose that there are three groups/conditions A, B, and C, and were interested in B vs A, and C vs B. Then, the proper contrast matrix is

versus	A	B	C
A	-	0	0
B	1	-	0
C	0	1	-

  The convention here is that the first group in the comparison is listed on the rows, and the second group is listed on the columns. If one specifies entry 1 for comparison of A vs B and that of B vs A simultaneously (redundant specification), then only one of two comparisons will be performed. Notice that specifying 1 on the lower and upper diagonal determines the direction of change (signs in the $\log_2$ fold change) in the analysis output. If all entries are zero, then mapDIA will perform peptide/protein quantification only, without the differential expression analysis.

- If module data is used, the following options need to be set:
  - **MODULE**: The name of the file containing the module data. Exclude this line when no protein-protein interaction data will be used.
  - **MODULE_TYPE**: The type of module information. Use "edge_list" if the module information file contains a list of edges (two columns of protein names separated by tabs, each row representing one "interaction" in interaction networks). Use "group_list" if the module information file is a list of protein/peptide groups (e.g. Gene Ontology Annotation File (GAF) Format 2.0). Exclude this line if no PPI interaction data is used.
  - **MRF_TYPE**: Use "-1_1" to indicate the model with penalty (default) or "0_1" to indicate the model with no penalty. Exclude this line if no PPI interaction data is used.
  - **MODULE_SIZE**: Two numbers separated by a whitespace, specifying the range of the group sizes to be considered in the MRF model. The first number is the smallest group size and the second number is the largest group size to be included. Both numbers should be ≥ 2. Set the second number to INF to remove the limit on large groups. Exclude this line if not using GAF 2.0.
  - **MODULE_FREQ**: The number groups the pair of proteins must appear in to be considered as an edge in the module. This number should be ≥ 1. Exclude this line if not using GAF 2.0.

- If a second set of module data is used, the following options need to be set:
  - **MODULE2**: as stated in **MODULE** for the second module data.
  - **MODULE_TYPE2**: as stated in **MODULE_TYPE** for the second module data.
  - **MRF_TYPE2**: as stated in **MRF_TYPE** for the second module data.
  - **MODULE_SIZE2**: as stated in **MODULE_SIZE** for the second module data.
  - **MODULE_FREQ2**: as stated in **MODULE_FREQ** for the second module data.

3 Input Data Formatting

As indicated above, there are two major experimental designs in the mapDIA software. In both cases for fragment-level data, the first line is expected to be the header, and the first three columns should be protein names, peptide names (if applicable), and fragment names (if applicable). Following these columns, the samples should be provided in a consistent manner as the input parameter file above.

For the IS design, the samples should be listed by groups, where the groups should appear in the same order as specified in the "LABEL". Within each group, however, the ordering of sample does not matter.

In the REP design, the conditions being compared should appear in the same order as specified in the "LABEL," and within the biological replicates should appear in the same order across all conditions. For example, if the design is a time course experiment with 3 time points (t1,t2,t3) across 2 biological replicates (A,B), then the conditions are time points and thus the samples should be organized in the following order (t1-A, t1-B) (t2-A, t2-B) (t3-A, t3-B).

See the two example input parameter files for the IS and REP designs in glycoproteomic data and 14-3-3 interactome data respectively, along with the supplemental table containing the input data file.

Important Remark. If RT δ option is chosen for normalization, the last column of the input file must contain retention time (synchronized across all samples) for each fragment (or peptide or protein).

4 Execution of the software

When the input data and input parameter file is ready, use the following command line call in the working directory:

```
> ./mapDIA/mapDIA  input_parameter_filename_here
```

for Linux/macOS or

```
> .\mapDIA\mapDIA_win64.exe  input_parameter_filename_here
```

for the Windows executable.

5 Analysis Output Table

The command line above will produce multiple output files, each carrying the information regarding fragment filtering/selection and post-processing input data, model fitting, and the score table (main output).

5.1 log2_data.txt

This file contains log_2-transformed fragment-level data after all the fragment selection (filtering) was applied as requested by the user. This file will therefore contain a smaller number of fragments than the original input data.

5.2 param.txt

This file contains the statistical model fitting, which shows the proportion of DE proteins in the list of all proteins after the fragment selection step. This can serve as a measure of goodness of fit if there is a prior expectation about how many proteins should be differentially expressed.

5.3 fragment_selection.txt

This table reports detailed information of fragment selection steps for each fragment, regarding whether each fragment was removed, and at which step it was removed. This output serves as an important indicator of how much the data will be reduced as a result of the fragment selection steps and can be used to tune the optimal level of filtering. Note that the column for SDF indicates whether a fragment had any values removed by SDF due to point-wise deviation from the median value.

5.4 analysis_output.txt, analysis_output_wide_format.txt

This table is the main output of the mapDIA software. The included fields are

- Protein: Protein ID
- nPeptide: the number of peptides used for modeling
- nFragment: the number of fragments used for modeling
- Label: group numbers being compared
- Label2: group labels being compared

- log2FC: $\log_2$ fold change between the groups or conditions
- log2FC_SE: standard error of the $\log_2$ fold change
- score: Probability score of differential expression. When module data is used, this column gives the score without incorporating the module data.
- FDR: The estimated Bayesian false discovery rate. When module data is used, this column gives the FDR without incorporating the module data.
- log_oddsDE: log(odds ratio of DE). When module data is used, this column gives the log odds ratio without incorporating the module data.

In the case of the REP design, there are extra columns providing each replicate specific summaries:

- log2FC1, ..., log2FCr, the $\log_2$ fold change for replicate r.
- nUp: the number of replicates with positive $\log_2$ fold change for the given comparison
- nDown: the number of replicates with positive $\log_2$ fold change for the given comparison

In the case where module data is used, there are extra columns providing analysis results incorporating module data

- score_: Probability score of differential expression for proteins forming the module (proteins with $\geq$1 links to other proteins).
- FDR_: The estimated Bayesian false discovery rate for proteins forming the module.
- log_oddsDE_: log(odds ratio of DE) for proteins forming the module.

5.5 fragments_for_protein_quantification.txt (fragment-level data only)

This file reports the intensities of fragments after normalisation and filtering. Filtered out fragments (printed with its IDs and "NA" for all its intensities) appears at the bottom of the table.

5.6 peptide_level.txt (fragment-level data only)

This file reports the intensities of fragments per peptide per sample summed up. The last column reports the number of fragments per peptide used in the analysis.

5.7 protein_level.txt

This file reports the intensities of fragments per protein per sample summed up. The last two column reports the number of fragments per protein and the number of peptides per protein used in the analysis.

5.8 duplicates.txt

This file reports the row names that appear more than once in the data.

6 Copyrights

Copyright (C) <2015-2017> Guoshou Teo < guoshou@u.nus.edu > and Hyungwon Choi < hyung_won_choi@nuhs.edu.sg >, National University of Singapore.

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.