Normalization and Quality Control of LC-MS Metabolomics Data

Leonid Brodsky[1,2,*], Arieh Moussaieff[1], Nir Shahaf[1], Asaph Aharoni[1] and Ilana Rogachev[1]

[1]Department of Plant Sciences, Weizmann Institute of Science, P.O. Box 26, Rehovot 76100, Israel

[2]Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

Figure S-1 . (**A**) An advantage of quantile normalization over no-normalization (or linear one) in between-replicate correlations of one sample against samples of the same replicate group. The group of samples is CL from the Arabidopsis-Four-Tissues dataset. The Pearson correlation coefficients between CL_cl2 sample and all other samples of the CL group are depicted. (**B**) An advantage of quantile normalization over linear one in between-replicate Euclidean distances of a sample against samples of the same replicate group for the case of CL_cl2 sample against all other samples of the CL group is depicted.

Figure S-2. PCA plots of sample distributions before and after quantile normalization. The circles depict groups of biological replicates; triangles depict the associated groups of technical controls. **(A)** The initial PCA distribution of biological and technical samples (five biological groups and five technical control groups: see **B** for index). **(B)** The distribution of biological and technical samples after quantile normalization of two "super groups"**;** all biological samples are normalized as one group and all technical controls are normalized as another group. **(C)** The profiles of log-intensities of mass-peaks after normalization of two "super-groups" (all biological samples and all technical controls).

Figure S-3. The Zcorr profiles across XCMS outputs from 50 different parameter settings applied to the analysis of three LC-MS data sets: (i) samples of tomato introgression lines (tomato ILs) with regions derived from the wild tomato (Lycopersicon pennellii) in chromosomes 1 and 2 (tomIL_chr1-2); (ii) samples of tomato ILs with regions of the wild tomato in chromosomes 3 and 4 (tomIL_chr3-4); (iii) whole roots samples derived from the Arabidopsis ecotypes. Arrow **B** depicts the best Zcorr parameter setting obtained for the ILs samples (R36 for both) while arrow **A** shows the best Zcorr parameter setting (R14) obtained for the ecotypes data set.

Figure **S-4**. The problematic fragments in the RT-m/z peak sorting from the original Arabidopsis-Four-Tissues data set: The figure represents two detected fragments from the RT-m/z peak sorting that are enriched by discrepancies between the profiles of two sample replicates [cl4 (dark blue) and cl5 (pink)]. The scaled down total fragment standard deviation (SD) score of each fragment is represented by dashed, yellow bars.

Figure **S-5**. The distribution of problematic fragments over RT-m/z plane in the original Arabidopsis-Four-Tissues dataset. The template of peak positions for this dataset contains 15,000 peaks. Similarly to the Zcorr best XCMS output (Fig. 5), the RT zones 1 – 5 min and 15 – 18 min are enriched by problematic fragments, but there is no concentration of discrepancies in the RT interval 12.5 – 14 min. In comparison with Fig.5, the smaller number of problematic fragments here is because the overall fitting of replicates in the original Arabidopsis-Four-Tissues dataset is lower than the fitting in the best XCMS output (see Fig. 2). Hence, the overall standard deviation (SD) of inter-replicate deviates is higher, and the segmentation method finds fewer intervals enriched by the inter-replicate deviates. Nevertheless, despite the absence of the RT zone 12.5 – 14 min, other zones of high inter-replicate deviations are in corroboration with Fig.5. Such a corroboration of two problematic RT zones between outputs of two different peak-picking algorithms may give evidence that inter-replicate deviations in these zones originates either from the biological samples or from the experimental procedure itself.

Figure **S-6**. Normalized total score of problematic fragments across samples in two outputs of the Arabidopsis-Four-Tissues experiment. The distribution of per-peak normalized scores of local inter-replicate-deviation intervals across samples for two analysis sets: the best XCMS parameter set and the original Arabidopsis-Four-Tissues data set 21. The graph shows that the two biological sample groups, cauline leaves (CL) and rosette leaves (RL) are enriched with inter-replicate deviations since the profiles of these groups across samples are relatively high and correlated. This provides additional evidence that the majority of inter-replicate deviations stem from either the biological samples or from the experimental procedure and not from wrong peak-picking or peak-alignment procedures. (ST - stem tissue; FL – inflorescence).

Figure S-7. Distribution of problematic fragments over the RT-m/z plane for the “ground truth” alignment of chromatograms. The template of peak positions on the RT-m/z plane for analysis of local discrepancies in the M1 and M2 data sets[19]. The M1 and M2 data sets contain 1,007 and 2,015 peaks, respectively. Several peak positions in (A) and (B) are highlighted by colors that indicate the scores of problematic fragments between pairs of replicates that cover this particular peak. **(A)** Peak positions on a template of aligned chromatograms of the M1 data set. A number of peaks are colored to denote the problematic fragments. The RT region at 3.2 min. is enriched with significant problematic intervals. **(B)** Peak positions on a template of aligned chromatograms of the M2 data set. Almost no peaks are colored to denote the problematic positions. **(C)** Problematic intervals of the M1 data set across RT. **(D)** Problematic intervals of the M2 data set across RT. There are no defective intervals of significant scores here.

Program S-1 (Program’s executable)

Program S-2 (Program instructions)

Program S-3 (working example of the multiInput mode)

Program S-4 (working example of the singleInput mode)