Journal of Chromatography A 1620 (2020) 460999

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

QPMASS: A parallel peak alignment and quantification software for the analysis of large-scale gas chromatography-mass spectrometry (GC-MS)-based metabolomics datasets

Lixin Duan[a,b,1], Aimin Ma[a,c,d,1], Xianbin Meng[a], Guo-an Shen[e], Xiaoquan Qi[a,d,*]

[a] Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[b] International Institute for Translational Chinese Medicine, Guangzhou University of Chinese Medicine, Guangzhou 510006, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100049, China
[e] Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China

ARTICLE INFO

Article history:
Received 3 November 2019
Revised 24 February 2020
Accepted 24 February 2020
Available online 26 February 2020

Keywords:
QPMASS
GC-MS
Metabolomics
Data analysis
Parallel computing

ABSTRACT

Gas chromatography-mass spectrometry (GC-MS) is a robust analytical platform for analysis of small molecules. Recently, it is widely used for large-scale metabolomics studies, in which hundreds or even thousands of samples are analyzed simultaneously, producing a very large and complex GC-MS datasets. A number of software are currently available for processing GC-MS data, but it is too compute-intensive for them to efficiently and accurately align chromatographic peaks from thousands of samples. Here, we report a newly developed software, QPMASS, for analysis of large-scale GC-MS data. The parallel computing with an advanced dynamic programming approach is implemented in QPMASS to align peaks from multiple samples based on retention time and mass spectra, enabling fast processing large-scale datasets. Furthermore, the missing value filtering and backfilling are introduced into the program, greatly reducing false positive and false negative errors to be less than 5%. We demonstrated that it took only 8 h to align and quantify a GC-TOF-MS dataset from 300 rice leaves samples, and 17 h to process a GC-qMS dataset from 1000 rice seed samples by using a personal computer (3.70 GHz CPU, 16 GB of memory and > 100 GB hard disk). QPMASS is written in C++ programming language, and is able to run under Windows operation system with a user-friendly interface.

1. Introduction

Metabolomics analysis of genetic population or natural population of crop cultivars are effective in identification of metabolic quantitative trait loci (mQTLs) that control metabolite contents and their related agronomical traits [1–4]. Typically, hundreds or even thousands of biological samples are analyzed in these mGWAS (metabolic genome-wide association studies) and mQTLs studies. Since gas chromatography-mass spectrometry (GC-MS) is a robust analytical platform with higher sensitivity and resolution, it has been widely applied in metabolomics analysis [5]. GC-MS analysis typically generates many fragment ions for each analytic compound, which makes the tasks of sample deconvolution and peak alignment very challenging. Thus the processing of GC-MS data is generally great sophisticated and time-consuming. So far, many software have been developed for analysis of GC-MS data (Table 1). Some software like AMDIS [6,7], MetaboliteDetector [8], ADAP [9–12], PyMS [13], MS-DIAL [14], and ChromaTOF (LECO, St. Joesph MI, USA) were developed for peak deconvolution. Among them, AMDIS is widely used for the deconvolution of GC-quadrupole MS (GC-qMS) data, while ChromaTOF is more often used for deconvolution of gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) data.

Another challenging issue in GC-MS analysis is about retention time drift. To solve this issue, some different strategies are applied in various software such as ADAP [9–12], PyMS [13], MS-DIAL [14], XCMS [15,16], MathDAMP [17], TagFinder [18], TargetSearch [19], MetAlign [20,21], flagme [22], and ChromAlignNet [23]. The nonlinear retention time alignment method (the observed deviation usually changes over time in a nonlinear mode within a sample, and these changes are fitted using a local polynomials regression fitting method-loess) is used to correct the retention time drift in XCMS [15]. Dynamic time warping (DTW) with an explicitly specified time shift is used for the alignment of datasets in MathDAMP,

* Corresponding authors.
E-mail address: xqi@ibcas.ac.cn (X. Qi).
[1] These authors contributed equally to this article.

https://doi.org/10.1016/j.chroma.2020.460999
0021-9673/

Table 1
Summary of software for analyzing GC-MS data.

No.	Name	Algorithm	Import data format	Description	Citation times
1	XCMS (CLI) [15,16]	Non-linear retention time alignment is used to correct for retention time drift	NetCDF; mzXML;MzData; mzML	Large-scale GC-MS data analysis, but it producing highly redundant datasets	2488△
2	AMDIS (GUI) [6,7]	The model peak method is used for peak deconvolution; signal to noise values is used to distinguish signal from noise at low signal levels	Almost all of GC-MS data format	Spectra deconvolution and metabolites identification, but peak alignment seems not be performed	695△
3	MetAlign (CLI) [20,21]	Two modes of alignment (rough and iterative alignment) are used	Masslynx .raw; mzData; Xcalibur .raw; NetCDF;mzXML; Agilent .d	Data pre-processing and peak alignment; the maximum number of files that can be processed in one session is 1000	492△
4	TagFinder (GUI) [18]	Linear interpolation between retention time anchors to calculate retention index; Pearson/ Spearman correlation is applied to find correlated clusters of tags	MetAlign output;NetCDF	Alignment of large GC-MS data into data matrix, but peak smooth and baseline correction are not available, and software download is not available currently as the link to the download page is missing	335
5	MCR (GUI) [38]	Batch modeling is used to study the dynamic behavior of the resolved metabolites over time	NetCDF; CSV	Processes all samples simultaneously and identify mass spectra of overlapping peaks, but it is sensitive to the number of co-analyzed files	309
6	MS-DIAL (GUI) [14]	MS^2Dec algorithm based peak deconvolution, least squares optimisation is used to extract model peaks from chromatograms	Analysis Base File (ABF) format; MzML; AB Sciex (.Wiff); Thermo Fisher Scientific (.RAW); Agilent Technologies (.D); Waters (.RAW); Bruker Daltonics (.D)	Peak deconvolution, peak alignment and compound identification	288
7	MET-IDEA (CLI) [39,40]	A calculated fixed value correction or a linear correction is applied to correct retention time	AMDIS .elu; NetCDF	Target metabolome analysis, an input list of Ion-retention time pair (IRt) list is required	198△
8	Metabolite Detector (GUI and CIL) [8]	A five point cubic Savitzky-Golay filter is applied to smooth spectral; deconvolution is performed using an improved algorithms applied by Colby et al. and Stein; a Gaussian function is used for the RI based similarity index calculation; chromatographs are aligned by retention time	NetCDF; JEOL FastFlight2	Data deconvolution and alignment, but it relies on QT4 based graphical user interface to ensure compatibility of cross platform	196
9	MathDAMP (CLI) [17]	Dynamic time warping (DTW) with an explicitly specified time shift function is used to align data	ChemStation .ms;mzXML; netCDF; Analyst .csv	Visualization and identification of differences between complex metabolite profiles, but the current release lacks of quantification function	112
10	TargetSerach (GUI and CIL) [19]	Peak apex intensities are used for peak picking; retention time index (RI) is used to retention time alignment	NetCDF	Data pre-processing and metabolites identification	96
11	ADAP (CLI) [9–12]	*k*-medoids clustering analyses (ADAP 1.0), model peak approach (ADAP 2.0 and 3.0) and multivariate curve resolution (ADAP-4.0) are applied to deconvolute coeluting components; extracted ion chromatograms (EIC) based two-phase approach is used for peak alignment	NetCDF	Peak deconvolution and alignment; it represents the solution to metabolites co-eluting analysis.	88△
12	MetaQuant (GUI) [41]	A mixture of internal standards with known concentrations are used to solve the integrals of the corresponding peaks in the spectra, then peak areas and the known substance concentrations are used for regression analysis; retention indices is used to minimize retention time shift	NetCDF; CSV	Target metabolome analysis, accurate quantification of GC-MS data, but a compound library is required	44
13	ChromaTOF (GUI) LECO, St. Joseph, MI, USA	Without published algorithm descriptions	Leco file formats	Spectra deconvolution and metabolites identification, it is developed especially for the dataset from its own GC-TOF-MS instrument	38*
14	PyMS (CLI) [13]	Moving-average and Savitzky-Golay filters are used for noise filter; dynamic programming is used for peak alignment	ANDI-MS/NetCDF; JCAMP-DX	Peak deconvolution and quantitation; dynamic programming based peak alignment; Message Passing Interface (MPI) based parallel data processing	31

(*continued on next page*)

Table 1 (*continued*)

No.	Name	Algorithm	Import data format	Description	Citation times
15	ChromAli gnNet (CLI) [23]	Pairwise comparisons and hierarchical clustering algorithm is used to group peaks	csv	Peak alignment	0
16	flagme (CLI) [22]	A similarity matrix is calculated based on a scoring function; dynamic programming is used for peak alignment	AMDIS .elu; NetCDF	Dynamic programming-based alignment strategy and data visualization	NP

GUI: Graphical user interface; CLI: Command line interface.
Citation times are obtained from "Web of Science" (http://isiknowledge.com) at the time of writing (2019/10/29) by searching correspongding references or key words (*); △: the citation times are the sum of the citations of corresponding references; NP: the corresponding reference was not found.
#: stands for import data formats of AMDIS including NetCDF, Agilent files, Bruker files, Agilent MS Engine files, Finnigan GCQ files, Finnigan INCOS files, Finnigan ITDS files, INFICON files, Micromass files, JEOL/Shrader file, Kratos Mach3 file, Xcalibur Raw file, MassLynx NT file, Shrader/GCMate file, Shimadzu MS files, PerkinElmer files, Varian MS files, Varian XMS file, Varian SMS file.

in which a representative set of peaks is used to align all datasets [17]. The peak alignment is performed also based on a reference peak list in MS-DIAL [14]. The retention time of internal substances is used to calculate retention index (RI) using linear interpolation between retention time anchors in TagFinder. The calculated RI and mass spectra are used to sort mass fragments, and then the mass fragments of same compound across all samples are binned and aligned into the mass tags, which are grouped into different time groups using the RI slide windows. Pearson/Spearman correlation is applied to find correlated clusters of tags in the time groups [18]. In TargetSearch, RI is used to align peaks against the retention time [19]. The rough and iterative alignment are applied in MetAlign. In the rough mode, mass peaks are grouped within a user defined time window, which then slide through the time dimension of all datasets to be aligned; the iterative alignment mode uses the same algorithm as that in the rough mode, and it applies the iterative calculation of retention time difference regard to the landmarks (the peaks that present in all datasets are chosen as landmarks). The iterative calculation will stop when the sliding time window is in the order of a mass peak width [20]. Two-phase approach based on extracted ion chromatograms (EIC) for peak alignment is provided in ADAP. In phase 1, mass spectra corresponding to the same component among all the samples are identified. The phase 2 is the refinement of phase 1, and it is used to find the best representative spectra across all the samples [9]. In ChromAlignNet, all pairs of qualified peaks are firstly compared to get a set of groups, and then hierarchical clustering is implemented for group assignment [23]. Dynamic programming method that is widely used for the alignment of DNA sequences [24] is also proposed for alignment of LC-MS data [25] and GC-MS data [26,27]. In this approach, the similarity of each pairs of peaks are calculated based on the retention time and mass spectra. The global alignment is built progressively, starting with two most similar peak lists in the similarity tree derived from the pair-wise alignments. In the case of more than one experiments with multiple replicates are conducted, the replicate experiments are aligned firstly ("within-state alignment"), and then the within-state alignments are aligned ("between-state alignment") progressively [26]. This approach is employed in flagme [22], and PyMS [13]. Although these software have their own special algorithms, it still inconvenient for them to analyze large-scale datasets, considering the complexity, computational costs and misalignment rate [26,28,29]. For example, it took nearly 91 h for MetAlign to align 940 GC-TOF-MS samples [20] and the alignment accuracy of MetAlign in aligning standard compounds mixture was 74% [29].

The search results on “Web of Science” database showed that XCMS and MetAlign were the most frequently used software to process large-scale GC-MS data (Table 1). However, XCMS and MetAlign were originally developed for analysis of LC-MS data [26,28]. Being oriented toward the detection of single ion peaks, these software tend to over-interpret GC-MS data [13,26]. There have been some attempts to assemble these individual ion signals into fragmentation characteristic mass spectra, but this process is widely considered to be extremely challenging [18,30]. Meanwhile, the maximum number of files that could be processed in one session of MetAlign was 1000 [20]. The available computing power could be a bottleneck for XCMS to process a large amount of datasets [13].

Selection of the best quantitative ion (quant ion) is a critical step in GC-MS data analysis, and there are many strategies for quant ions selection in existing software. For example, ADAP and MetaboliteDetector are tend to choose those most intense and unique (not shared with neighbor peaks) ion as quant ions [8,9]. In PyMS, a single ion shared in all peaks in a certain time from N abundant ions in the alignment results is chosen [13]. In flagme, the intensity of all ions corresponding to a peak are calculated [22]. Peak quantification could sometimes be based on different quant ions for the same peak in different samples, making concatenation and comparison of metabolite levels between samples difficult. Taken together, it has become apparent that there are critical limitations of available algorithms and software tools for the analysis of ultra large metabolomics datasets, and it is necessary to develop a new software suitable for efficiently analyzing both GC-qMS and GC-TOF-MS data.

Here we report a newly developed program, QPMASS, for researchers in multiple disciplines to easily analyze their ultra-large GC-MS datasets for mGWAS, mQTLs and many other studies. In this software, the parallel computing with an advanced dynamic programming approach, a three-parameter strategy for selection of optimal quant ions, as well as missing value filtering and backfilling are implemented to rapidly and accurately alignment and quantification of large-scale datasets. QPMASS is written in C++ programming language, and able to run under commonly used Windows operation system. It accepts the popular raw data format of netCDF or mzXML from GC-TOF-MS and GC-qMS, as well as the peak deconvolution results from AMDIS or ChromaTOF.

2. Materials and methods

2.1. Chemicals and reagents

Methanol, chloroform and water (HPLC grade) were purchased from Fisher Scientific (Hampton, NH). Pyridine, N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) reagent, methoxylamine hydrochloride, the internal standard compounds of adonitol and nonadecanoic acid, and the authentic standard compounds were all purchased from Sigma-Aldrich.

2.2. Samples preparation

The authentic standard compounds were prepared as 2 mM stock solutions (dissolved in HPLC grade water), except L-tyrosine

and L-serine were prepared as 400 μM stock solutions. Mixtures of these compounds at certain concentrations were then prepared as detailed in Supplementary Table 1 [31] (adonitol was used as internal standard [32–34] and its concentration was kept constant in all samples). These mixtures were referred to as the artificial sample group A and B (Asa and Asb). Each group contained six samples. Equal volume of each sample were mixed as the quality control (QC) sample. 100 μL of the mixed QC sample, three blank samples (100 μL HPLC grade water with same concentration of adonitol as that in other samples was set as blank sample), and twelve artificial samples (Asa and Asb) were dried in a rotary vacuum evaporator without heating. The dried residues were oximated with 40 μL methoxylamine hydrochloride (20 mg/mL) in anhydrous pyridine at 37 °C for 2 h, and then silylated at 37 °C for 30 min with 70 μL MSTFA. The derivatized samples were then transferred into 250 μL glass vials (Agilent) for GC-TOF-MS analysis.

Leaves of rice cultivars [35] were harvested at 60 days after sowing. Each sample had two biological replicates. The lyophilized leaves were ground into powder with a mixer mill (Retsch Mixer Mill MM 400). Metabolites extraction was prepared as previously described [36]. 1 mL methanol/chloroform/water = 2.5:1:1 (v/v/v) with internal standard of adonitol was added to 20 mg of powder. The same volume of extracting solution and internal standard was used as blank sample. Metabolites were extracted in a shaker (220 rpm) at 37 °C for two hours, followed by centrifugation at 12,000 rpm for 10 min. 700 μL of supernatant was transferred into a new tube. 1 mL of methanol/chloroform = 1:1 (v/v) (nonadecanoic acid as internal standard) was used to suspend the precipitate, and followed by extracting in a shaker (220 rpm) at 37 °C for one hour. After being centrifuged (12,000 rpm, 10 min), 700 μL supernatant was transferred into above new tube. 350 μL distilled water was added, which was used to separate chloroform phase from methanol/water phase. The sample was centrifuged for 2 min at 5,000 rpm. Finally, 200 μL methanol/water phase and 200 μL chloroform phase were transferred into 250 μL glass vials (Agilent), separately. Both methanol/water phase and chloroform phase were then dried and dissolved in 50 μL methoxylamine hydrochloride (20 mg/mL, pyridine), and incubated at 37 °C for two hours with continuous shaking. Then 80 μL MSTFA was added, and the samples were incubated at 37 °C for 30 min with continuous shaking. Derivatized samples were stored at room temperature for two hours before GC-TOF-MS analysis to ensure complete derivation.

The lyophilized seeds of 500 rice *japonica* varieties were ground into powder with a mixer mill. 100 mg powder were weighted, and each sample had two biological replicates. Metabolites extraction was the same as above, then the extracts were analyzed by GC-qMS.

2.3. GC-TOF-MS and GC-qMS analysis

The GC-TOF-MS system was composed of an Agilent 6890 (Agilent Corporation, USA) gas chromatograph and a LECO Pegasus IV time-of-flight mass spectrometer (LECO Corporation, USA). 1 μL of each derivatized sample was injected into the gas chromatograph equipped with a 30 m × 0.25 mm I.D. fused silica capillary column with a chemically bonded 0.25 μm DB-5 MS stationary phase (J&W ScientiWc, USA). The injector temperature was at 280 °C. The flow rate of helium gas through the column was 1 mL/min. The column temperature was held at 60 °C for 1 min, and then elevated to 220 °C at a rate of 5 °C/min, and held there for 2 min. The temperature was then elevated to 310 °C at a rate of 15 °C/min, and held there for 2 min. The transfer line and ion source temperatures were at 280 °C and 200 °C, respectively. Ions were generated by a 70 eV electron beam, and 20 spectra per second were recorded in the 50–600 m/z mass range. The acceleration voltage was turned on after a solvent delay of 300 s. The detector voltage was 1670 V.

The samples of rice seeds were analyzed using an Agilent 5977A MS coupled to Agilent 7890B GC instrument under electronic impact at 70 eV with an Agilent DB-5 HT column (30 m × 0.32 mm × 0.1 μm). The oven temperature was initially set at 70 °C, and then was raised from 70 °C to 290 °C (10 °C/min), and maintained there for 4 min. The temperature was then elevated from 290 °C to 310 °C (20 °C/min), and hold there for 10 min. The injection volume was 1 μL.

2.4. Data deconvolution and export

The raw data was deconvoluted by ChromaTOF using the following settings: the baseline was set to 1, the smoothing threshold was set to 7, and the signal-to-noise value was set to 10. The exported .csv files of deconvolution results contained peak name, retention time, quant mass, peak area, and relative mass spectra (Supplementary Table 2). The .csv and .cdf files were imported into QPMASS for alignment and quantification. The parameters used in QPMASS were setting as follows: mZMin = 50, mZMax = 700, deconvolution data type = .csv, top_n = 20, slaveCounts = 1, raw data type = netCDF, gap = 0.08, cutoff(s) = 6, peakthreshold = 0.15, mzthreshold = 0.01, backfill = 1, dtRange(s) = 0.1, simThreshold = 0.4, areaTol = 0.15. The Automated Mass Spectral Deconvolution and Identification System (AMDIS, Version 2.62, NIST, US) was also used to deconvolute peaks, but we did not use the library matching function of AMDIS. While running AMDIS software, “Shape requirements” was set to “Medium”, “Sensitivity” was also set to “Medium”, the “Type of analysis” parameter was set to “Simple”, “Adjacent Peak Subtraction” was set to “One”, and the “Component width” was set to 32. The obtained .elu result files were aligned using the R package “flagme”. The key alignment parameters used in flagme were set as following: wn.gap = 0.5, wn.D = 0.05, bw.gap = 0.06, and bw.D = 0.02. The quantitative accuracy of QPMASS was compared with XCMS and ChromaTOF. In XCMS, the ion peak detection method was “centWave”, “S/N value” was set to 50, “peakwidth” was set to 3–10, and the “nearest method” was used for grouping peaks.

3. Theory

In order to shorten the time for processing large-scale datasets, the parallel computing with an advanced dynamic programming approach, which is based on Robinson's dynamic programming algorithm research [26] is implemented in QPMASS to align peaks from multiple samples. To reduce both false positive and false negative errors, the missing value filtering and backfilling are introduced into QPMASS. Moreover, a three-parameter strategy is developed for the selection of optimal quant ions for peak quantification.

The workflow of QPMASS includes seven main steps (Fig. 1). (1) Files import. QPMASS requires both deconvoluted peak data and raw data, which can be submitted through the deconvolution data path and mzXML path/cdf path, respectively. QPMASS will process peak firstly if the raw mzXML or netCDF files are used, whereas the deconvolution data can be directly used for aligning peaks. (2) Sample distribution. All samples are divided into subsets according to a preliminary hierarchical clustering analysis. (3) Peak alignment. QPMASS implements a parallel computing method for aligning peaks. The alignment results (aligned mass spectra) can be used to identify metabolites using the NIST spectral search engineer/database. (4) Raw data pre-processing. The raw data is pre-processed prior to peak integration through baseline correction, smoothing, and denoising. (5) Quantitative ion selection. The alignment results are then used in the selection of quant ions for peak quantification. Quant ions are selected based on three key data characteristics, including peak shape, peak-to-peak

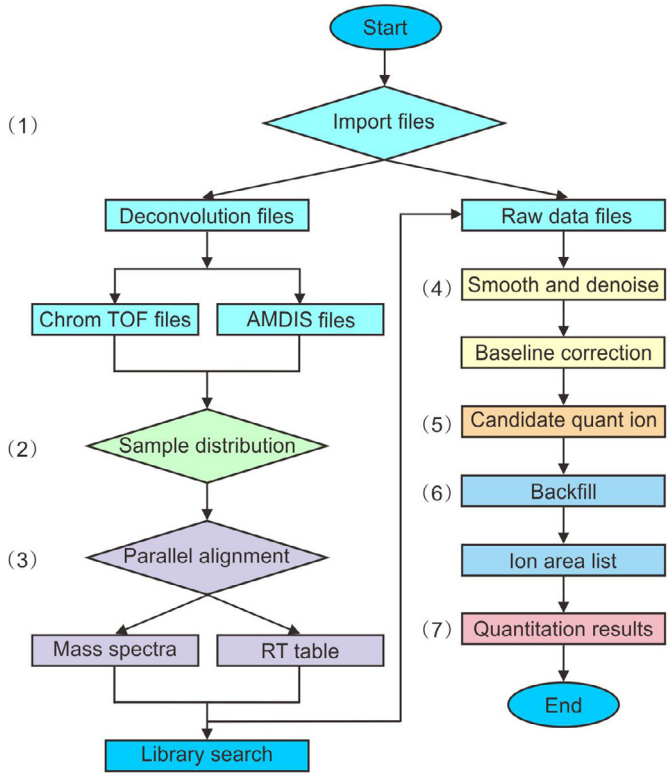

Fig. 1. Workflow of QPMASS for processing GC-MS data. Arabic numbers stand for the processing steps of QPMASS.

separation, and ion intensity. (6) Missing value filtering and backfilling. QPMASS can backfill (integrate) peak area values for each of missing data in each sample. (7) Files export. QPMASS mainly exports the aligned mass spectra, the aligned retention time, and the quantified peak area data.

4. Results

4.1. Sample distribution and parallel peak alignment

Prior to peak alignment, the samples are split into subsets that will be processed by different computer threads. In QPMASS, the samples are split into subgroup according to the similarity assessment results from a furthest-neighbor joining clustering based hierarchical clustering method, in which samples are grouped into the allotted number of subsets based on their closeness of clustering. With this method, the number of samples in each subset corresponds to the total number of samples are divided by the number of available processing threads.

In QPMASS, the peak alignment is essentially based on Robinson's dynamic programming algorithm research [26] (Fig. 2A). The peak similarity function *P(i,j)* (Eq. (1)) gives the similarity score between the mass spectra of peaks *i* and *j*. *S(i,j)* is calculated as the cosine of angle between the vectors of two mass spectra (Eq. (2)). The retention time tolerance parameter *"D"* can be modulated using the selected weight given to retention time in total similarity matching. A gap (match-to-nothing) is defined as a missing value in an alignment. Peaks that cannot be aligned (corresponding to a gap) are designated with a value of "NA" (missing value). The gap penalty is set as a parameter "*G*". It is possible to deploy dynamic programming to find the global solution and achieve an optimal alignment of all peaks. *W(i,j)* (Eq. (3)) is used to find the global solution (i.e., to align) of two alignments such as M-alignment and N-alignment, where *I* is the indicator function and *P* is given by Eq. (1). Finally, the score matrices (Eq. (3)) and the trace back resulting from the application of dynamic programming are used to deduce the optimal alignment.

The parallel alignment approach implemented in QPMASS greatly accelerates the processing speed. In QPMASS, parallel peak alignment starts from aligning samples within each subset, followed by aligning consensus samples (subset reference) from different subsets, which is derived from the average mass spectra and retention time in each subset. The final alignment result is a combination of all subsets based on the alignment results of consensus samples. For example, six samples were divided into two subsets according to hierarchical clustering analysis. S1, S2 and S3 were in subset1, the others in subset2. In the alignment procedure, the peaks in each subset were aligned separately, then the consensus peaks for each subset (subset1 reference, and subset2 reference) were produced based on the average mass spectra and retention time of all the samples in each subset. The alignment result of all consensus samples, which was global reference, was used to combine the subset results in the final alignment results (Fig. 2B).

4.2. Quantitative ion selection

To ensure accurate quantitation of peaks in different samples and give a better basis for quant ion selection, we developed a three-parameter strategy for the selection of optimal quant ions. According to this strategy, the peak-to-peak separation, peak shape, and ion intensity were all taken into consideration. In contrast, the ion with highest intensity in a well separated peak is chosen as quant ion in other software.

The three parameters used for selecting of quant ions include "ion intensity", "*height_ratio*", and "*sum_ratio*". Fig. 3A shows a simulation of an extracted-ion chromatogram for two peaks, which is used to illustrate the principle of quant ion selection in QPMASS. For ion intensity, QPMASS sets a parameter "top_n" for n number of the highest intensity ions from an aligned mass spectra. Using *m/z* 73 as an example, *m* represents the position of peak apex; *r* and *l* are the right and left borders of the peak, respectively; *H* is the intensity of the ion at the apex; h_r is the intensity of the ion at the right border; h_l is the intensity of the ion at the left border; and h_i is the intensity of ion at some point between *l* and *r*. "*height_ratio*" is the ratio of the intensities of an ion at the right and left borders to the intensity of ion at the apex, which is used to assess the overlapping situation between adjacent peaks (Eq. (4)). The larger the "*height_ratio*" is, the greater the degree of peak overlapping. "*sum_ratio*" is the average ratio of the intensity of an ion at each scan in the peak interval to the intensity of the ion at the apex (Eq. (5)), representing the degree of sharpness of peak shape; the smaller this value is, the sharper the peak is. An optimal quant ion should have a smaller *height_ratio* firstly, then a smaller *sum_ratio*. If an optimal quant ion cannot be found by assessing the first two parameters, the ion with the highest intensity will be chosen.

4.3. Missing value filtering and backfilling

AMDIS and ChromaTOF often yield many false positive deconvolution results [37]. Some of these specious, false positive peaks are only present in a small number of samples, and cannot be aligned in most samples. In QPMASS, a user defined filtering parameter called "peakThreshold" is used to filter these peaks. For example, if the "peakThreshold" is set to 80%, peaks that are not present in 80% of the samples (in a processing subset) are removed from the alignment results. The filtering parameter of "peakThreshold" is only applied at the level of a subset in the parallel alignment routine. After filtering the specified percentage of putative false positive peaks using the "peakThreshold" parameter, there typically still exists large number of missing values

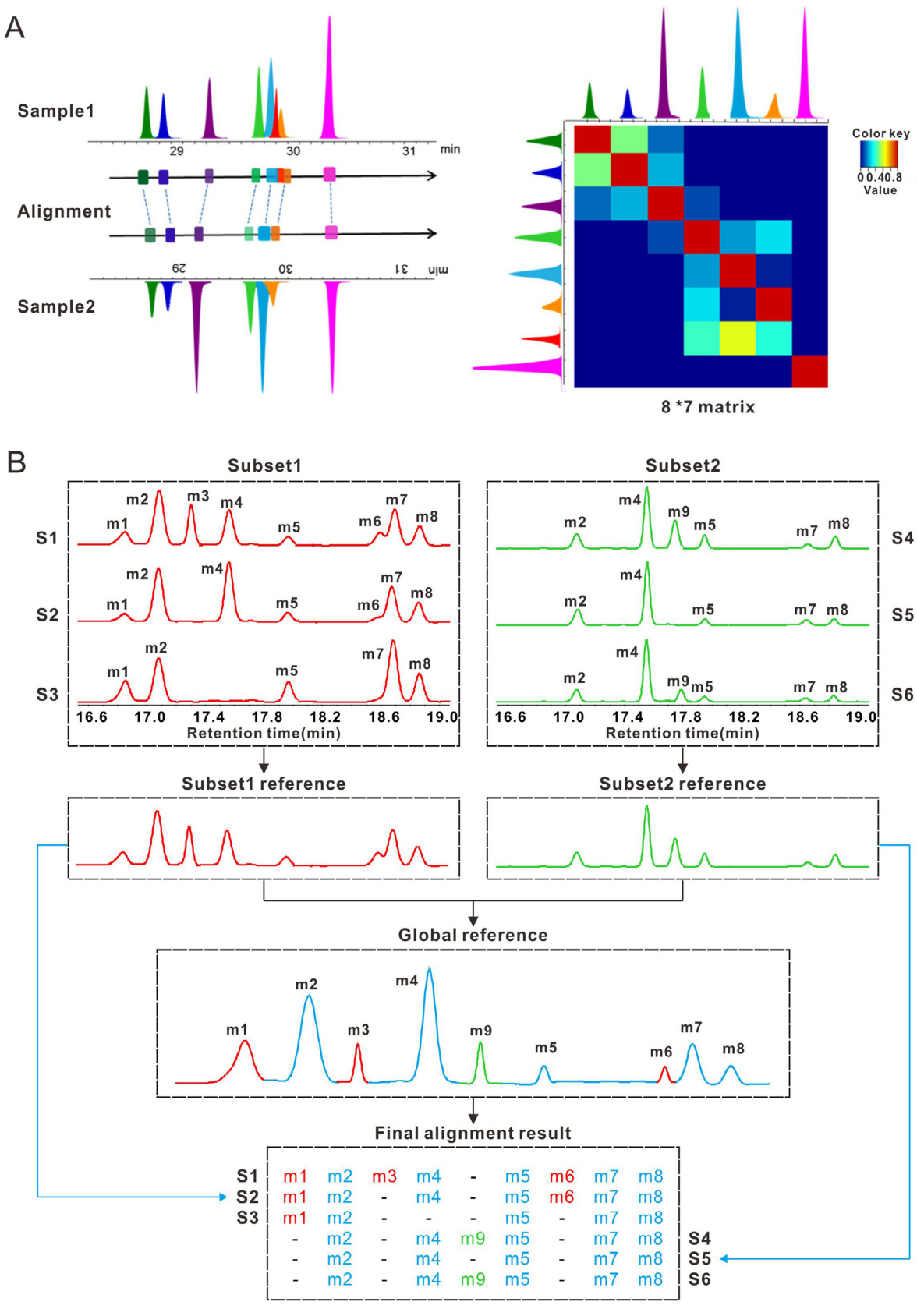

Fig. 2. A diagram of using dynamic programming and parallel peak alignment in QPMASS. (A) Diagram of using dynamic programming for peak alignment. The left panel shows the matched peaks after peak alignment. There are eight and seven peaks detected in sample 1 and 2, respectively. Except the red peak in sample 1, others can be aligned within two samples. The right panel is the score matrix for the peak alignment of two samples. The legend color corresponds to the similarity value of mass spectra, and the two axes refer to the detected peaks in sample 1 and 2. (B) A diagram of parallel peak alignment in QPMASS. S1-S6 represent six samples, and m1-m9 are the detected peaks in these samples. Subset1 reference and subset2 reference are the consensus samples derived from the average mass spectra and retention time of all the samples in subset1 and subset2, respectively. Global reference is the alignment result of these two consensus samples (subset1 reference and subset2 reference). The final alignment result is the combination of each subset alignment result based on the global reference result. Peaks and characters in red indicate peaks that only detected in subset1; in green mean peaks only exited in subset2; and in blue represent the peaks shared by both two subsets. Dash line means the peak was missing in the corresponding sample.

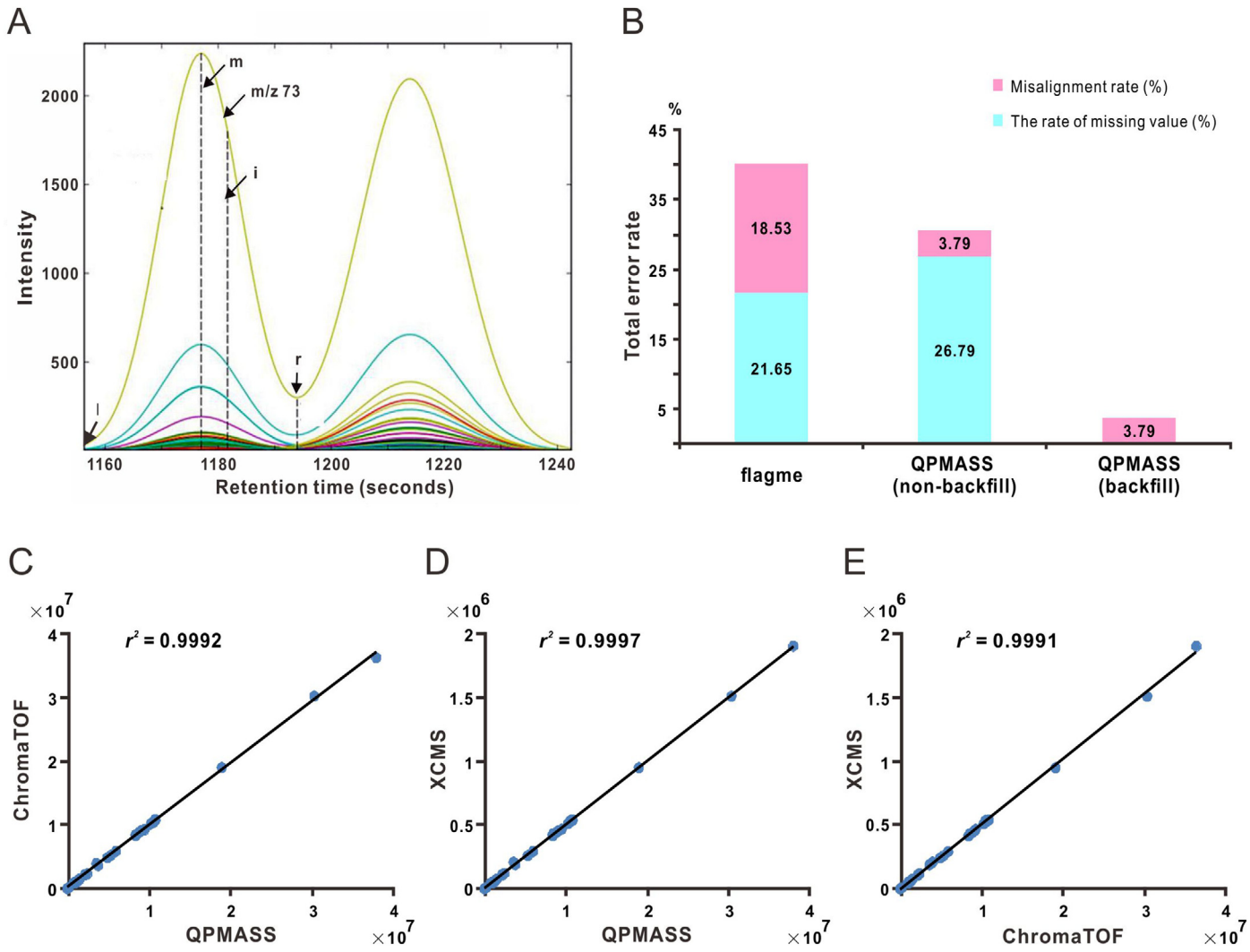

Fig. 3. The alignment and quantification ability of QPMASS. (A) The three-parameter strategy for selection of the quantitative ions. Lines in different colors represent ions that are selected as the top "*n*" ions with high intensity from the alignment results. "*m*" is the position of the peak apex; "*r*" and "*l*" are the right and left borders of the peak; "*i*" is a specified point of the peak. (B) Comparison of alignment accuracy between QPMASS and flagme. Peaks detected in twelve artificial samples and four QC samples were aligned in this procedure. "total error rate" is the sum of misalignment rate and the rate of missing value. (C-E) The correlation of peak areas obtained from QPMASS, XCMS and ChromaTOF. The average abundance of each peaks among four QC replicates were used to compare the quantification performance of the above three software.

in the alignment results (false negative), which might actually correspond to peaks that are genuinely present in some samples. Backfilling of missing data is considered to be necessary for robust statistical analysis [15]. and it has been used in some software, like XCMS, in which the quant ion from the aligned results is used to retrieve and integrate the area for missing peaks [15]. In QPMASS, for missing peaks, mass spectra of the quant ion is used to search peaks with high similarity within defined range of the retention time, and then the area of the missing peak is estimated, so that both quantitative backfill and qualitative judgment can be ensured. In this procedure, the parameter "simThreshold" is used to define the similarity of the mass spectra. QPMASS could export either backfilled or non-backfilled quantification results. The combination of filtering and backfilling can partially eliminate false positives, and fill in the majority of missing values.

4.4. Assessment of alignment accuracy and quantification performance

To validate the alignment and quantification accuracy of QPMASS, two groups of artificial samples (ASa and ASb), representing two types of biological samples, were prepared by mixing twenty-six standard compounds (Supplementary Table 1). Equal volume of each of the ASa and ASb samples were mixed to generate a “quality control” (QC) sample that was the representative of whole metabolite pool. This QC sample was analyzed for four times with GC-TOF-MS. Additionally, manual extraction of exact retention time and mass spectra for each constituent compounds of ASa and ASb samples were determined by analyzing each compound individually with GC-TOF-MS.

We assessed various existing software, and found it difficult to directly compare QPMASS with them. Since AMDIS, ADAP, MS-DIAL are specialized in peak deconvolution [6,9,12,14]. MS-DIAL and MathDAMP require a reference library [14,17]. MathDAMP is lack of quantification function [17]. TagFinder and TargetSearch need retention time index information [18,19]. ChromAlignNet may highly demand of computer memory [23]. MCR is sensitive to co-analyzed files [38]. Furthermore, a few of them are mainly for target metabolome analysis [39–41]. And some software are currently not available as the link to the download page is missing [13,18,20]. So we mainly compared the alignment performance of QPMASS with flagme, because flagme used the same dynamic programming method for peak alignment. XCMS is a popular software for single ion alignment, and has a good quantitation perfor-

mance, which has been widely used in GC-MS analysis [42–44]. ChromaTOF is another widely used quantitation software for GC-TOF-MS data. So we compared XCMS (R version) and ChromaTOF with QPMASS to evaluate the quantitation accuracy of QPMASS.

We used “total error rate” to evaluate the alignment performance of QPMASS and flagme. The term of “total error rate” is defined as the ratio of number of missing and misaligned peaks to the total number of detected peaks. Too many missing values is a serious problem in the GC-MS data alignment process, which strongly affects the subsequent multivariate statistical analysis. ADMIS and ChromaTOF often deconvolute a peak into multiple similar mass spectra components. In our results, the ion signals for the reference standard compound of glycine in ASa1 (Supplementary Fig. 1A) and ASa2 (Supplementary Fig. 1B) were deconvoluted by ChromaTOF into four components, respectively (Supplementary Fig. 1C–1J). Components N153 (Supplementary Fig. 1D) and N155 (Supplementary Fig. 1F) in the ASa1 sample and components N148 (Supplementary Fig. 1H) and N149 (Supplementary Fig. 1I) in the ASa2 sample all had very similar mass spectra and very close retention time. Although components N153 and N155 were both from glycine, they were aligned as two individual peaks. Likewise, the glycine deconvolution components N148 and N149 in sample ASa2 were also aligned as two separate peaks. In other samples in the ASa series (e.g. ASa3, ASa4), glycine was deconvoluted as a single peak. This discrepancy leads to a situation where ASa3 and ASa4 had missing values (filled by "NA") for the "doubled" glycine peaks of ASa1 and ASa2 (Supplementary Table 3). It is difficult for most of the software to discriminate this situation. We used backfilling approach to re-integrate the missing values ("NA"). When QPMASS alignment was performed without backfilling, the total error rate was 30.58%, while the total error rate for flagme was 40.18%. After backfilling by using QPMASS, the total error rate was dramatically reduced to 3.79% (Fig. 3B, Supplementary Data Sets 1), while the process of backfilling is not available in flagme. Multiple deconvoluted peaks that correspond to a single compound will have the same area values and can thus be detected easily in the result tables (Supplementary Table 3). Specious components such as component N147 (Supplementary Fig. 1G) in the ASa2 sample had high noise signals and could not be aligned with other samples. These peaks could be filtered with the “peakThreshold” parameter.

In some cases, some components with similar mass spectra but large different in retention time, might be incorrectly aligned as the same peak. QPMASS used a peak retention time threshold parameter "cutoff(s)" to prevent the alignment of distant peaks. “cutoff(s)” values are typically 6-10 seconds. The values for this parameter should be chosen based on user’s knowledge of the retention time shift and instrumental conditions. Misalignment errors are most likely to occur in sugars like glucose, fructose and galactose (Supplementary Data Sets 1). Such sugars generally generate multiple derivatization peaks, and their mass spectra are very similar to those of other sugars. In our test, the misalignment error rate of flagme was 18.53%, whereas the misalignment error rate of QPMASS was 3.79% (Fig. 3B). This reduction in the error rate resulted from using the "cutoff(s)" filter in QPMASS. However, the percentage of "NA" values in QPMASS was greater than that in flagme as a result of using this filter parameter (Fig. 3B).

The quantification performance of QPMASS was compared with XCMS (R version) and ChromaTOF. Since it was difficult to select the same quantitative ions that used in flagme, so we could not directly compare the quantitative accuracy with flagme. We addressed this discrepancy by manually specifying which ion peak to use in ChromaTOF and XCMS. Except for the peaks with misalignment errors, the peak areas from the QPMASS peak integration all showed good correlation ($r^2 > 0.99$) with the peak areas from ChromaTOF and XCMS (Fig. 3C–3E, Supplementary Data Sets 2).

4.5. Assessment of computing speed for processing large-scale datasets

During our test period, flagme failed to complete the processing of the above mentioned artificial samples under Windows operation system due to limited computing resources. Thus we switched to a powerful Linux system (a dual-core CPU server with 251 GB of memory). As a result, flagme still spent one hour and 37 min to process these samples. Similarly, a much longer time was needed for ChromaTOF. By contrast, QPMASS completed it in few minutes in a personal computer (Intel Xeon E3-1230 V2 with 16 GB of memory). So in the assessment speed of QPMASS in processing metabolomics data, we mainly compared with XCMS (R version) using data of GC-TOF-MS (LECO) and GC-qMS (Agilent) from different number of plant samples. For GC-TOF-MS data, all the samples were deconvoluted by ChromaTOF. The .cdf files were imported into XCMS for quantification and alignment, while both .csv and .cdf files were used to align and quantify peaks in QPMASS. The task of aligning 50 samples using XCMS took nearly 212 min under an Intel Xeon E3-1230 V2 with 16 GB of memory, while QPMASS could finish this task in only six minutes. When the number of samples increased to 200, XCMS needed almost 18 h to align these samples, whereas QPMASS only needed nearly 2 hours (Fig. 4A). Furthermore, it only needed nearly 17 hours for QPMASS to process GC-qMS data of 1000 samples, whereas XCMS only processed about 400 samples within the same time (Fig. 4B).

5. Discussion

So far, the existing software for analyzing GC-MS data are quite limited in their capability to process large numbers of samples. XCMS is one of the most frequently used software for metabolite profiling. In our study, it took almost 18 hours to analyze 200 GC-TOF-MS samples using XCMS (Fig. 4A). With the increasing of the number of samples, the processing time increased exponentially, making the analysis of very large datasets untenable. The newly developed QPMASS software, is designed with the analysis of very large datasets in mind, which acquires significant advantages in both speed and capability to handle large-scale GC-MS data from different sources. QPMASS achieves ideal results by employing parallel computing approach to align multiple samples, three-parameter strategy for identification of suitable quantitative ions for accurate quantification, and the missing value filtering and backfilling to reduce the alignment errors.

The major challenge for aligning peaks is how to deal with false positive peaks and missing values. Any error introduced during previous peak detection will further worsen the alignment step later [26]. False positive peaks in alignment results are frequently related to the chemical contaminates and various noise signals. Large numbers of false positive peaks will cause severe problems during peak alignment, yielding unexpected results with many missing values. The number of false positive peaks needs to be eliminated in some other way, such as manual inspection of data. QPMASS can filter missing values automatically and thus reduce the number of specious peaks and retain the peaks present in the majority of samples. The combination of filtering and backfilling can partially eliminate false positives and fill in the majority of missing values. Although QPMASS used a state-of-the-art algorithm for peak alignment, there was still 3.79 % misalignment error rate in our tested examples (Fig. 3B). We checked these errors out individually, and found that most of mismatches occurred in the alignment of peaks came from sugars such as glucose, lyxose, galactose and fructose (Supplementary Data Sets 1). In contrast, the amino acid and fatty acid peaks were markedly less prone to have alignment errors. The main reason for this discrepancy is that the structures of sugars are quite similar, as well as their

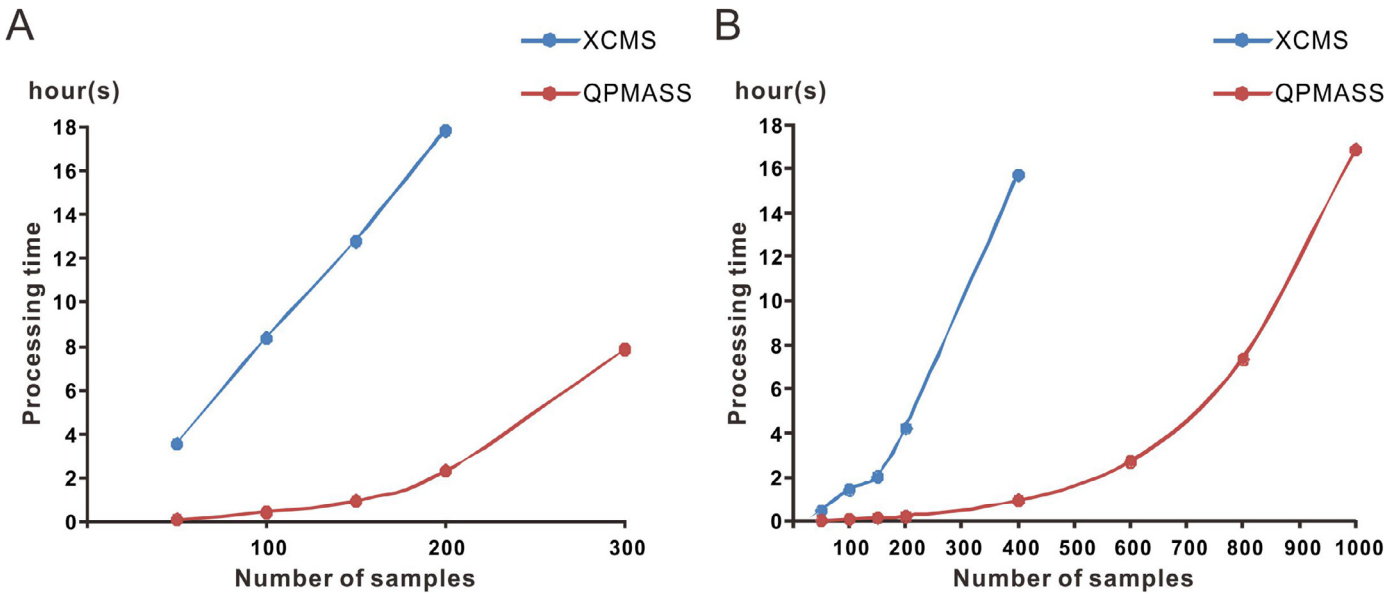

Fig. 4. The comparison of processing time of QPMASS and XCMS. (A) The processing speed for processing GC-TOF-MS data by using XCMS and QPMASS. (B) The processing speed for processing GC-qMS data by using XCMS and QPMASS. Lines in blue indicate the processing time by using XCMS; lines in red indicate the processing time by using QPMASS.

retention time and mass spectra. To solve this problem, we used methoxiamine in pyridine to treat samples in the first step of two-step derivatization method. This will beneficially keep sugars in open conformations in order to minimize the total number of conformational states and relieve steric hindrances for silylation, but there are still multiple derivative peaks with this method [45]. It is difficult for any peak alignment algorithm to distinguish these highly similar peaks. One possible solution is to control the derivatization conditions and GC conditions to reduce or separate these peaks. Using standards to compare the retention time and mass spectra is another good solution.

In QPMASS, before peak alignment, samples were divided into subsets according to their similarity. We had used it to process a large number of samples in mQTLs and mGWAS studies, and we did not find substantial misalignment in our result, which may be due to that highly genetically related samples were used in our studies. However, there will have some problems sometimes, especially when a few samples differ greatly from others. In this case, grouping samples based on different condition (experiment design) may achieve better alignment. We will introduce a new module of artificial grouping in the next version of QPMASS.

6. Conclusions

The newly developed QPMASS is specially designed for peak alignment and quantification for large-scale GC-TOF-MS and GC-qMS datasets, which show high computing performance and can rapidly and automatically process large-scale GC-MS metabolomic datasets used in mQTL and mGWAS studies. For brief manual of how to use QPMASS, please see the user manual. QPMASS is available for academic and non-commercial use at ftp://download.big.ac.cn/QPmass/QPmass_V1.0.zip.

Declaration of Competing Interest

None

CRediT authorship contribution statement

Lixin Duan: Conceptualization, Investigation, Methodology, Data curation, Writing - original draft, Writing - review & editing. **Aimin Ma:** Investigation, Methodology, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Xianbin Meng:** Methodology, Data curation. **Guo-an Shen:** Data curation, Writing - review & editing. **Xiaoquan Qi:** Conceptualization, Investigation, Project administration, Writing - original draft, Writing - review & editing.

Acknowledgments

We thank Hongjun Ren for the technical assistance. We thank Bin Han, Lu Wang from Plant Science Facility of the Institute of Botany, Chinese Academy of Sciences for their excellent technical assistance on GC-TOF-MS and GC-qMS. This research was supported by the National Key Research and Development Program of China (2016YFD0100904), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB27010202), the National Natural Science Foundation of China (31530050, 81874333, 31570306) and Science and Technology Program of Guangzhou, China (2018-1002-SF-0437).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2020.460999.

Appendices

$$P(i,j) = S(i,j) \cdot \exp\left[-\frac{(t_i - t_j)^2}{2D^2}\right] \quad (1)$$

$$S(A,B) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

$$W(i,j) = \frac{\sum_{b=1}^{M}\sum_{a=1}^{N} P(p_{ia}, q_{jb})}{\sum_{b=1}^{M}\sum_{a=1}^{N} I\left[P(p_{ia}, q_{jb}) > 0\right]} \quad (3)$$

$$height\_ratio = \frac{h_r}{H} + \frac{h_l}{H} \quad (4)$$

$$sum\_ratio = \frac{\sum_{i=m}^{r} \frac{h_i}{H} + \sum_{i=l}^{m} \frac{h_i}{H}}{r - l} \quad (5)$$

References

[1] J.J.B. Keurentjes, J.Y. Fu, C.H.R. de Vos, A. Lommen, R.D. Hall, R.J. Bino, L.H.W. van der Plas, R.C. Jansen, D. Vreugdenhil, M. Koornneef, The genetics of plant metabolism, Nat. Genet. 38 (2006) 842–849, doi:10.1038/ng1815.

[2] J. Krumsiek, K. Suhre, A.M. Evans, M.W. Mitchell, R.P. Mohney, M.V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski, C. Gieger, F.J. Theis, G. Kastenmüller, Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information, PLoS Genet. 8 (2012) e1003005, doi:10.1371/journal.pgen.1003005.

[3] X.L. Gong, W. Chen, Y.Q. Gao, X.Q. Liu, H.Y. Zhang, C.G. Xu, S.B. Yu, Q.F. Zhang, J. Luo, Gentic analysis of the metabolome exemplified using a rice population, Proc. Nalt. Acad. Sci. USA 110 (2013) 20320–20325, doi:10.1073/pnas.1319681110.

[4] S. Alseekh, T. Tohge, R. Wendenberg, F. Scossa, N. Omranian, J. Li, S. Kleessen, P. Giavalisco, T. Pleban, B. Mueller-Roeber, D. Zamir, Z. Nikoloski, A.R. Ferniea, Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato, Plant Cell 27 (2015) 485–512, doi:10.1105/tpc.114.132266.

[5] J. Luo, Metabolite-based genome-wide association studies in plants, Curr. Opin. Plant Biol. 24 (2015) 31–38, doi:10.1016/j.pbi.2015.01.006.

[6] J.M. Halket, A. Przyborowska, S.E. Stein, W.G. Mallard, S. Down, R.A. Chalmers, Deconvolution gas chromatography mass spectrometry of urinary organic acids-potential for pattern recognition and automated identification of metabolic disorders, Rapid Commun. Mass Spectrom. 13 (1999) 279–284, doi:10.1002/(SICI)1097-0231(19990228)13:4⟨279::AID-RCM478⟩3.0.CO;2-I.

[7] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J. Am. Soc. Mass Spectrom. 10 (1999) 770–781, doi:10.1016/S1044-0305(99)00047-1.

[8] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, MetaboliteDetector, comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis, Anal. Chem. 81 (2009) 3429–3439, doi:10.1021/ac802689c.

[9] W.X. Jiang, Y.P. Qiu, Y. Ni, M.M. Su, W. Jia, X.X. Du, An automated data analysis pipeline for GC-TOF-MS metabonomics studies, J. Proteome Res. 9 (2010) 5974–5981, doi:10.1021/pr1007703.

[10] Y. Ni, Y.P. Qiu, W.X. Jiang, K. Suttlemyre, M.M. Su, W.C. Zhang, W. Jia, X.X. Du, ADAP-GC 2, ADAP-GC 2.0: deconvolution of coeluting metabolites from GC-TOF-MS data for metabolomics studies, Anal. Chem. 84 (2012) 6619–6629, doi:10.1021/ac300898h.

[11] Y. Ni, M.M. Su, Y.P. Qiu, W. Jia, X.X. Du, ADAP-GC 3, ADAP-GC 3.0: improved peak detection and deconvolution of co-eluting metabolites from GC/TOF-MS data for metabolomics studies, Anal. Chem. 88 (2016) 8802–8811, doi:10.1021/acs.analchem.6b02222.

[12] A. Smirnov, Y.P. Qiu, W. Jia, D.I. Walker, D.P. Jones, X.X. Du, ADAP-GC 4, ADAP-GC 4.0: application of clustering-assisted multivariate curve resolution to spectral deconvolution of gas chromatography-mass spectrometry metabolomics data, Anal. Chem. 91 (2019) 9069–9077, doi:10.1021/acs.analchem.9b01424.

[13] S. O'Callaghan, D.P. De Souza, A. Isaac, Q. Wang, L. Hodkinson, M. Olshansky, T. Erwin, B. Appelbe, D.L. Tull, U. Roessner, A. Bacic, M.J. McConville, V.A. Likić, PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools, BMC Bioinform. 13 (2012) 115, doi:10.1186/1471-2105-13-115.

[14] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, M. Arita, MS-DIAL, data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 (2015) 523–526, doi:10.1038/nmeth.3393.

[15] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS, processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787, doi:10.1021/ac051437y.

[16] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS Online, XCMS online: a web-based platform to process untargeted metabolomic data, Anal. Chem. 84 (2012) 5035–5039, doi:10.1021/ac300698c.

[17] R. Baran, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert, M. Tomita, MathDAMP, a package for differential analysis of metabolite profiles, BMC Bioinform. 7 (2006) 530, doi:10.1186/1471-2105-7-530.

[18] A. Luedemann, K. Strassburg, A. Erban, J. Kopka, TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments, Bioinformatics 24 (2008) 732–737, doi:10.1093/bioinformatics/btn023.

[19] A. Cuadros-Inostroza, C. Caldana, H. Redestig, M. Kusano, J. Lisec, H. Peña-Cortés, L. Willmitzer, M.A. Hannah, TargetSearch-a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data, BMC Bioinform. 10 (2009) 428, doi:10.1186/1471-2105-10-428.

[20] A. Lommen, MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, Anal. Chem. 81 (2009) 3079–3086, doi:10.1021/ac900036d.

[21] A. Lommen, H.J. Kools, MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware, Metabolomics 8 (2012) 719–726, doi:10.1007/s11306-011-0369-1.

[22] M.D. Robinson, flagme: analysis of metabolomics GC/MS data, R package version 1.14.0 (2010).

[23] M. Li, X.R. Wang, Peak alignment of gas chromatography-mass spectrometry data with deep learning, J. Chromatogr. A 1604 (2019) 460476, doi:10.1016/j.chroma.2019.460476.

[24] S.R. Eddy, A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, BMC Bioinform. 3 (2002) 18, doi:10.1186/1471-2105-3-18.

[25] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data, J. Chromatogr. A 961 (2002) 237–244, doi:10.1016/S0021-9673(02)00588-5.

[26] M.D. Robinson, D.P. De Souza, W.W. Keen, E.C. Saunders, M.J. McConville, T.P. Speed, V.A. Likić, A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments, BMC Bioinform. 8 (2007) 419, doi:10.1186/1471-2105-8-419.

[27] H.H. Yang, H.J. Ren, L.Q. Li, L.X. Duan, T. Guo, L.L. Du, X.Q. Qi, Multiple samples alignment for GC-MS data in parallel on Sector/Sphere, J. Comput. Appl. 33 (2013) 215-218. https://doi.org/1001-9081(2013)33:1<215:JYSSDQ>2.0.TX;2-Z.

[28] Y. Koh, K.K. Pasikanti, C.W. Yap, E.C.Y. Chan, Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data, J. Chromatogr. A 1217 (2010) 8308–8316, doi:10.1016/j.chroma.2010.10.101.

[29] W.H. Niu, E. Knight, Q.Y. Xia, B.D. McGarvey, Comparative evaluation of eight software programs for alignment of gas chromatography-mass spectrometry chromatograms in metabolomics experiments, J. Chromatogr. A 1374 (2014) 199–206, doi:10.1016/j.chroma.2014.11.005.

[30] Y. Tikunov, A. Lommen, C.H.R. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles, Plant Physiol. 139 (2005) 1125–1137, doi:10.1104/pp.105.068130.

[31] L.X. Duan, I. Molnár, J.H. Snyder, G.A. Shen, X.Q. Qi, Discrimination and quantification of true biological signals in metabolomics analysis based on liquid chromatography-mass spectrometry, Mol. Plant 9 (2016) 1217–1220, doi:10.1016/j.molp.2016.05.009.

[32] M.J. Kim, M.Y. Lee, J.C. Shon, Y.S. Kwon, K.H. Liu, C.H. Lee, K.M. Ku, Untargeted and targeted metabolomics analyses of blackberries-Understanding postharvest red drupelet disorder, Food Chem. 300 (2019) 125169, doi:10.1016/j.foodchem.2019.125169.

[33] W. Zhang, H.B. He, X.D. Zhang, Determination of neutral sugars in soil by capillary gas chromatography after derivatization to aldononitrile acetates, Soil Biol. Biochem. 39 (2007) 2665–2669, doi:10.1016/j.soilbio.2007.04.003.

[34] B.B. Misra, E. de Armas, Z.H. Tong, S.X. Chen, Metabolomic responses of guard cells and mesophyll cells to bicarbonate, PLoS ONE 10 (2015) e0144206, doi:10.1371/journal.pone.0144206.

[35] H.R. Wang, X. Xu, F.G. Vieira, Y.H. Xiao, Z.K. Li, J. Wang, R. Nielsen, C.C. Chu, The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication, Mol. Plant 9 (2016) 975–985, doi:10.1016/j.molp.2016.04.018.

[36] W. Weckwerth, K. Wenzel, O. Fiehn, Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks, Proteomics 4 (2004) 78–83, doi:10.1002/pmic.200200500.

[37] H.M. Lu, Y.Z. Liang, W.B. Dunn, H.L. Shen, D.B. Kell, Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS, Trends Anal. Chem. 27 (2008) 215–227, doi:10.1016/j.trac.2007.11.004.

[38] P. Jonsson, A.I. Johansson, J. Gullberg, J. Trygg, J. A, B. Grung, S. Marklund, M. Sjöström, H. Antti, T. Moritz, High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses, Anal. Chem. 77 (2005) 5635–5642, doi:10.1021/ac050601e.

[39] C.D. Broeckling, I.R. Reddy, A.L. Duran, X.C. Zhao, L.W. Sumner, MET-IDEA: data extraction tool for mass spectrometry-based metabolomics, Anal. Chem. 78 (2006) 4334–4341, doi:10.1021/ac0521596.

[40] Z.T. Lei, H.Q. Li, J. Chang, P.X. Zhao, L.W. Sumner, MET-IDEA version 2.06, improved efficiency and additional functions for mass spectrometry-based metabolomics data, Metabolomics 8 (2012) S105–S110, doi:10.1007/s11306-012-0397-5.

[41] B. Bunk, M. Kucklick, R. Jonas, R. Münch, M. Schobert, D. Jahn, K. Hiller, MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data, Bioinformatics 22 (2006) 2962–2965, doi:10.1093/bioinformatics/btl526.

[42] Y.M. Zhang, Y.Y. Zhang, Q. Zhang, Y. Lv, T. Sun, L. Han, C.C. Bai, Y.J. Yu, Automatic peak detection coupled with multivariate curve resolution-alternating least squares for peak resolution in gas chromatography-mass spectrometry, J. Chromatogr. A 1601 (2019) 300–309, doi:10.1016/j.chroma.2019.04.065.

[43] Y.C. Feng, T.X. Fu, L.Y. Zhang, C.Y. Wang, D.J. Zhang, Research on differential metabolites in distinction of rice (*Oryza sativa* L.) origin based on GC-MS, J. Chem. (2019) 1614504, doi:10.1155/2019/1614504.

[44] R. Fernandez-Varela, G. Tomasi, J.H. Christensen, An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes, J. Chromatogr. A 1384 (2015) 133–141, doi:10.1016/j.chroma.2015.01.025.

[45] F.G. Xu, L. Zou, C.N. Ong, Multiorigination of chromatographic peaks in derivatized GC/MS metabolomics: a confounder that influences metabolic pathway interpretation, J. Proteome Res. 8 (2009) 5657–5665, doi:10.1021/pr900738b.