BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2636–2643
doi:10.1093/bioinformatics/btu359

Data and text mining

Advance Access publication May 28, 2014

Massifquant: open-source Kalman filter-based XC-MS isotope trace feature detection

Christopher J. Conley[1,*], Rob Smith[2], Ralf J. O. Torgrip[3], Ryan M. Taylor[4], Ralf Tautenhahn[5,6,7] and John T. Prince[4,*]

[1]Department of Statistics, University of California Davis, Davis, CA 95616, [2]Department of Computer Science, Brigham Young University, Provo, UT 84606, USA, [3]Department of Analytical Chemistry, Stockholm University, SE-106 91, Stockholm, Sweden, [4]Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84606, [5]Department of Chemistry, [6]Department of Molecular Biology and [7]Center for Metabolomics, The Scripps Research Institute, La Jolla, CA 92037, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Isotope trace (IT) detection is a fundamental step for liquid or gas chromatography mass spectrometry (XC-MS) data analysis that faces a multitude of technical challenges on complex samples. The Kalman filter (KF) application to IT detection addresses some of these challenges; it discriminates closely eluting ITs in the m/z dimension, flexibly handles heteroscedastic m/z variances and does not bin the m/z axis. Yet, the behavior of this KF application has not been fully characterized, as no cost-free open-source implementation exists and incomplete evaluation standards for IT detection persist.

Results: Massifquant is an open-source solution for KF IT detection that has been subjected to novel and rigorous methods of performance evaluation. The presented evaluation with accompanying annotations and optimization guide sets a new standard for comparative IT detection. Compared with centWave, matchedFilter and MZMine2—alternative IT detection engines—Massifquant detected more true ITs in a real LC-MS complex sample, especially low-intensity ITs. It also offers competitive specificity and equally effective quantitation accuracy.

Availability and implementation: Massifquant is integrated into *XCMS* with GPL license $\geq$ 2.0 and hosted by Bioconductor: http://bioconductor.org. Annotation data are archived at http://hdl.lib.byu.edu/1877/3232. Parameter optimization code and documentation is hosted at https://github.com/topherconley/optimize-it.

Contact: cjconley@ucdavis.edu or jtprince@chem.byu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 23, 2013; revised on May 19, 2014; accepted on May 21, 2014

1 INTRODUCTION

The most important automated data-analysis step in a typical quantitative -omics XC-MS analysis pipeline is the isotope trace (IT) detection (Cappadona *et al.*, 2012). In liquid or gas chromatography mass spectrometry (LC-MS or GC-MS, with either specified as XC-MS), analytes elute with chromatographic separation and are subsequently measured by the mass spectrometer. IT detection is the first and essential step in enumerating the signals of these analytes.

IT detection is a trivial task when performed on data derived from simple mixtures, but can be highly challenging for complex mixtures because there are (i) large numbers of analytes that co-elute, many show interlocking or overlapping isotope envelopes; (ii) an unknown number of analytes; (iii) an abundance of ITs with low signal-to-noise ratio; (iv) significant intensity variation in the signal composing lower abundance ITs because of dynamic range limitations of the spectrometer; and (v) heteroscedastic m/z variance as a function of intensity for most mass spectrometers. Unisotropic m/z variance means that the data comprising the tails of an IT have larger m/z variance than the data around the mode, and that low-abundance ITs have a larger m/z variance than high-abundance ITs.

Though difficult to achieve, increasing the sensitivity and accuracy of IT detection software influences the entire downstream analytical pipeline (Smith *et al.*, 2013a). An example: vast numbers of peptides go unidentified in proteomic analyses (Michalski *et al.*, 2011); a more sensitive IT detection would allow researchers to track and quantify these peptides, leveraging identifications acquired in other samples. It goes without saying that accurately determining IT boundaries and distinguishing signal from noise improves quantitation results. Furthermore, accuracy in IT detection can also result in more accurate precursor mass estimates and therefore yield an increase in both the number and quality of peptide identifications.

Most IT-detection software, such as matchedFilter, rely on the creation of fixed width m/z bins (buckets) to facilitate finding and quantifying eluting analytes. Though bucketing is computationally efficient, for complex datasets, it is impossible to find a bin size and position that excludes closely co-eluting ITs while also being broad enough to fully capture the IT of interest. To address this shortcoming, Tautenhahn *et al.* (2008) developed a software package, centWave, which uses a binless pre-scan to first identify regions of interest composed of centroids. A 'centroid' is a (m/z, intensity) measurement pair at a given time scan of the chromatographic dimension. Once a region is specified, the centroids are then collapsed into a one-dimensional chromatogram, and wavelet-based curve fitting is performed to separate closely eluting ITs. The approach is appealing because the

*To whom correspondence should be addressed.

initial algorithm identifies zones of interest in a binless way and because the algorithm used for detecting ITs using intensity fluctuation in the time domain is sophisticated. However, in this approach subtle shifts in m/z value are ignored when data are combined into a one-dimensional chromatogram. ITs that are close in m/z or with poor chromatographic profiles may not be properly resolved.

The same year Aberg *et al.* (2008) developed TracMass, which includes a binless IT detection algorithm that fully uses m/z information in distinguishing ITs. TracMass uses a chromatographically traversing 2D Kalman filter model (KF)—one dimension focused on m/z values and the other on intensity values—to determine which centroids belong with each extending IT. The decision to incorporate a centroid is made by carefully weighing all previous m/z and intensity evidence of that IT, so mis-incorporation of centroids is rare, as the KFs incorporate more data. Furthermore, the KF accounts for the heteroscedastic variance within the same IT as intensity values change. The KF approach can disentangle even the most closely eluting chromatographic ITs. Furthermore, for the non-expert user, TracMass requires few user parameters for effective operation.

Despite its apparent promise for IT detection in complex samples, no peer-reviewed publication had compared TracMass performance to leading options (Zhang *et al.*, 2009; Zhou *et al.*, 2012) until just recently with TracMass2 (Tengstrand *et al.*, 2014). This is not an isolated deficiency—most IT detection algorithms are not rigorously evaluated because of the difficulty of establishing ground-truth data, especially for lower abundance ITs (Zhou *et al.*, 2012; Smith *et al.*, 2013b). Other compelling binless methods for quantitation may benefit from a similar evaluation as presented here (Cox and Mann, 2008; Yu *et al.*, 2009).

Here, we make available an open-source implementation of the TracMass algorithm, called Massifquant, and integrate it into the popular XCMS software suite (Smith *et al.*, 2006; Tautenhahn *et al.*, 2008). Like TracMass, Massifquant uses a 2D KF to quickly, accurately and adaptively find ITs in highly complex samples without resorting to binning, and its open license (GPL $\geq$ 2.0) enables further extension and inspection. We indicate how the KF adapts to m/z variance and describe two major modifications, which mitigate known limitations of TracMass. We detail novel metrics for evaluating XC-MS IT detection and use these metrics with manually annotated data to perform a detailed evaluation of Massifquant, centWave, matchedFilter and MZMine2 (Pluskal *et al.*, 2010) performance on different LC-MS platforms.

2 METHODS

2.1 Description of the Massifquant algorithm

Massifquant relies on 2D KFs to identify ITs in XC-MS data. A single KF's purpose is to track the m/z and intensity coordinates of an IT over the chromatographic dimension. A 'track' is an instance of a KF model, which predicts the existence of a centroid in the next time scan. If the prediction is close enough to a real centroid, it incorporates the real centroid to the track. Closeness is determined by quasi-confidence intervals centered about the prediction. The KF then updates its estimate of the underlying 'true' centroid and predicts again. When the signal of the IT disappears (i.e. we have reached the end of a chromatographic IT), the KF will fail to predict a centroid on successive scans and tracking will be terminated.

With many ITs to be discovered, Massifquant manages a host of active KFs. For a given scan, each active KF claims the centroid that best fits its predicted location. Unclaimed centroids trigger new instances of KF tracks in the expectation that these are the beginning of new ITs. The process is then repeated on the next scan until all scans have been examined. In this way, every centroid is either claimed by an existing KF or triggers the creation of a new KF. After an entire sample has been parsed, spurious KFs are discarded based on simple filters for minimum length, intensity, expected m/z deviance or consecutive missed predictions.

We will describe the 'Kalman gain' to highlight the model's adaptive nature and how it can be tuned. After the KF predicts a centroid, it refines the prediction by carefully weighting the model prediction error through a modeling device known as the Kalman gain. This device is largely a function of (i) the estimation error covariance, which is initialized by the modeler, but evolves over time based on prediction performance; (ii) and the assumed measurement error of the mass spectrometer, also defined by the user. So the modeler may tune the Kalman gain based on these parameters. A smaller Kalman gain means that the model prediction, which is based on previous observations, is trusted to be closer to the true centroid location than the newly acquired observation. The default settings of Massifquant create a Kalman gain that places more trust in early acquired observations (i.e. the first 4–30 scans) as illustrated in Supplementary Figure S1. The idea is to find the IT's location quickly and to not deviate once it has been found; the default works for a variety of situations but can also be tuned to a particular dataset. The fact that the KF continuously adapts its centroid-prediction estimates based on the information it has previously amassed and the variance it encounters makes it an effective tool for identifying ITs with their own specific heteroscedastic variance. For a more mathematical discussion, an introduction to the theory behind the discrete Kalman filter/gain are described in Welch and Bishop (2006) and section 2 of the Supplementary Materials.

Massifquant implements most of the core of the TracMass algorithm; however, it is difficult to determine how much the two algorithms differ, as the latter's source code is not publicly available. There are a few known major differences. The initialization of the P is likely different. Moreover, the intensity component of the Scheffe-type quasi-confidence intervals—used to classify whether a next-scan centroid belongs to a KF prediction—was not found to be sufficiently discriminatory. Massifquant only uses the m/z dimension to determine a successful prediction. Retaining the intensity estimation in the KF does seem to aid in resolving competing KFs that claim the same centroid (by virtue of comparing their 2D prediction distances).

Massifquant also implements a function to ensure continuity of identified ITs that is not found in TracMass (discussed in section 3 of the Supplementary Information). We found that a KF will periodically lose the position of the IT, stop tracking it en route, triggering a new KF track that will finish estimating the IT's other data points (Supplementary Figure S2). As each KF track corresponds to an IT, we call the undesirable phenomenon 'segmentation'. The segmentation problem was addressed by an *ad hoc* t-test comparing the m/z locations between these problematic KF. The conservative test combines many of the segmented tracks into a unified IT.

A more thorough description of the Massifquant implementation is given in the Supplementary Material (see the section 'Reimplementing the Kalman filter model'). The Supplementary Materials highlight some differences with TracMass and a discussion of the logic behind specific design decisions. The description will be useful for anyone seeking to modify or extend the algorithm. Massifquant was written in C and has been integrated into the XCMS pipeline available through Bioconductor (Smith *et al.*, 2006; Gentleman *et al.*, 2004). It plays the same role as centWave, matchedFilter or MZMine2's IT detection algorithm in the differential analysis workflow.

2.2 Annotation

2.2.1 Datasets We chose two different LC-MS datasets to assess IT-detection flexibility. The first annotated dataset, MM14, is a subset from a UPLC-ESI-QTOF analysis of 14 plant metabolites resulting in 46 annotated ITs. The centWave developers originally showcased their method of parameter optimization on the entire dataset, and its provenance is detailed in Tautenhahn *et al*. (2008).

The second dataset, MOUSE, is one fraction from a larger mouse brain phosphoproteomic analysis. Briefly, 408.8 mg of brain tissue was homogenized/boiled in SDS-lysis buffer and clarified. Proteins were then digested and peptides purified following the filter-aided sample preparation (FASP) protocol (Wisniewski *et al*., 2009) to yield an estimated 7.3 mg of peptides. Titanspere TiO2 beads (25 mg; GL Sciences) were used to enrich for phosphorylated peptides. 3M Empore Anion Exchange disks were packed into a 200 l pipette, and Britton & Robinson buffer was used to elute at pH 11 (the fraction termed 'MOUSE' in this work), 6, 5, 4 and 2. MS analysis was performed with an LTQ-Orbitrap XL fed by an Eksigent NanoLC UHPLC system. A Nano Acquity (1.7 m, 130 C18 bead BEH, 75 mm × 150 mm) column run at 375 nl/min in a linear gradient from 2.5 to 10% acetonitrile (ACN) (with water and 0.1% formic acid as the second buffer) for 60 min, then to 28% ACN for an additional 220 min. The complete raw file is available on request, and virtually all parameters may be accessed using the cross-platform unfinnigan software (see https://code.google.com/p/unfinnigan/). The relevant parameters are MS1 data collected between 375–1800 m/z at 60 000 resolution with an MS/MS data-dependent scan collected after each MS scan. The section chosen for hand-annotation generally spans retention time of 5429.5–7306.2 s and 600.0003–637.3923 m/z. In total, this area contained 589 annotated ITs, which show variation in length, shape and variance.

2.2.2 Data annotation The MOUSE and MM14 datasets were manually annotated to be used as ground truth for assessing the automated IT-detection abilities. A tuned LC-nanoESI system is capable of producing consistent chromatographic IT shapes. However, when running complex samples, even on the best tuned system, fundamental dynamic range limitations will unavoidably produce IT shapes that are far from ideal. The lack of characteristic IT shapes among lower abundance ITs, the number of overlapping ITs (in m/z and time) and their sheer number and density makes manual annotation difficult. For the MOUSE data, any IT that did not exceed a maximum intensity of 1×10^5 was ignored to preserve the integrity of the annotation.

Because IT annotation in complex datasets is challenging, we established guidelines for what is called a true IT. These guidelines consider within-IT and between-IT characteristics to ensure the best annotation possible. To be defined as an IT, a series of centroids should typically exhibit the following properties:

Within

(1) The m/z error variance structure is influenced by intensity. Toward the tails of an IT, the m/z observations show mostly symmetric and increasing deviations from the mean. The body and apex centroids deviate less. From a bird's eye view (i.e. looking down the intensity axis), the m/z-time projection has the shape of a string fraying at the edges.

(2) The collective centroids should have a chromatographic IT shape. Dramatic oscillations in intensity from scan to scan could disqualify an annotation.

Between

(1) The detected ITs should have approximately the same m/z ppm variance.

(2) Within an isotopic envelope, ITs should have similar mode and shape, although length typically varies.

In each case, great effort was made to balance the benefits of the systematic application of these rules with human judgment. Each IT was individually annotated (based on all criteria) and then wrapped into appropriate isotopic distributions where possible.

We executed this annotation scheme on the MM14 and MOUSE datasets using Topp-View (Sturm and Kohlbacher, 2009) as follows: from a global 2D view, the annotator identified mass traces satisfying mentioned properties. After zooming, a 3D inspection confirmed similar chromatographic length and shape for a given isotopic distribution. Once confirmed, the IT's centroids were selected and collectively saved into an .mzML file. Candidate mass traces that did not sufficiently satisfy all the criteria, but still had some resemblance to an IT, were labeled as questionable and saved as .mzML files; these were excluded from the algorithm performance analysis, as they were deemed liable to interfere with true algorithmic specificity and sensitivity. Objectively determining an IT's chromatographic boundaries is difficult, especially because there is so much diversity among IT shape and length. Generally, we tried to include as much of each IT tail as possible and to be as consistent as possible across each dataset.

2.3 Performance evaluation

Different algorithms select different portions of an IT when attempting to identify ITs (any attempted IT classification we call a 'candidate'). Because the extent and location of the mapping from a candidate to the true IT may vary widely, gauging the success of a candidate can be challenging. For example, a method that identifies 30 centroids directly in the middle of the high-intensity region of an IT should be given more credit than one that identifies 35 centroids but that are all in the low-intensity tail region. In another example, credit should be given to an algorithm that successfully captures an entire IT with three distinct candidates, but it should not receive as much credit as an algorithm that identified the IT with a single candidate. These examples motivated the development of two ways of examining success: at the IT level and at the entire sample level.

2.3.1 Isotope trace-level evaluation Classifying the success of an algorithm at the IT level requires the classification to be general enough to handle a variety of IT shapes and yet to be precise. To classify the successful identification of an IT, we defined metrics that consider how a candidate's centroids individually contribute to the overall intensity of the annotated IT, namely, the true area under the curve (AUC). The centroids clustered into a candidate are either true-positive results, false-positive results or false-negative results. Restricting attention to the true-positive results, a candidate's true AUC is denoted as $\mathrm{AUC_{TP}}$. Naturally, a candidate's relative correct identification of an IT within the context of intensity is defined to be $\alpha := \frac{\mathrm{AUC}_{TP}}{\mathrm{AUC}_A}$. Now, an algorithm is said to sufficiently identify the ith annotated IT if $\alpha_i \geq 1 - r$, where $0 \leq r \leq 1$. For the following analysis, we took $r = 0.5$ because requiring a candidate to capture >50% of an IT's total intensity ensures that the main body of an IT has been identified, while still allowing for differences in opinion on exact IT boundaries. In short, this criterion abstracts away the difficulty of varying shapes and algorithmic-selection bias.

Conversely, the false-positive and false-negative centroids contain precise information as to where and by how much a candidate is accurate. To be clear, the AUC quantitation error is taking evaluation precision beyond classification. Let AUC* be the quantification reported by the algorithm, which includes true- and false-positive centroids alike and excludes false-negative centroids. Then, the AUC percent error is simply $\epsilon := \frac{|\mathrm{AUC}_A - \mathrm{AUC}^*|}{\mathrm{AUC}_A} \times 100\%$. Dramatic variation in IT intensity motivated the percent error representation.

Another issue is that true-negative ITs are impossible to define. So an algorithm's IT-identification accuracy was measured by the commonly used metrics of precision and recall (sensitivity) for information retrieval. *Isotope trace sensitivity* (s_I) is the number of ITs correctly identified by the

algorithm *divided by* the number of true ITs. *Isotope trace precision* (p_f) is the number of ITs correctly identified by the algorithm *divided by* the number of algorithm-claimed ITs. High sensitivity means the algorithm successfully identifies most true ITs, while high precision is a measure of identification reliability. The harmonic mean of these is the F_1 score $:=2\frac{s_f p_f}{s_f+p_f}$; it summarizes the overall identification performance.

2.3.2 Sample-level evaluation Finally, sample-level metrics allow us to define how much of the entire sample AUC was correctly identified without regard for individual ITs. It is a way to quantify the level of intensity information found by an IT detection without regard to how the centroids are actually clustered into ITs. *The sample sensitivity* is defined as $\frac{\sum_i \mathrm{AUC}_{TP_i}}{\sum_j \mathrm{AUC}_{A_j}}$. This is the total algorithm-identified true raw intensity divided by total true raw intensity. On this global level, a true negative can be defined as the sample noise, or the centroids that do not contribute to any real ITs. Thus, the *sample specificity* equals $\frac{\sum_i \mathrm{AUC}_{TN_i}}{\sum_j \mathrm{AUC}_{FP_j}+\sum_k \mathrm{AUC}_{TN_k}}$. This taken to be the total correctly algorithm-ignored raw intensity (true negative signal) divided by total noise raw intensity of the sample (including false positives of the algorithm). These last two metrics are useful as a global measure of accuracy in contrast to the IT-specific accuracy in the preceding metrics.

2.3.3 Evaluation by IT type An evaluation should indicate how certain IT types influence performance. Simpson's paradox further motivates an evaluation by type, as conclusions based on the aggregate annotation are sometimes reversed when analyzed by type (Bickel *et al.*, 1975). We classified ITs by intensity, ppm error and length. Annotated ITs were grouped by the variable of interest into 8 percentile categories $\{[0, 12.5\%), [12.5\%, 25\%) \ldots, [87.5\%, 100\%]\}$. For example, the longest IT was categorized in [87.5%, 100%]. The recall was computed for each category; precision was approximate because mapping the algorithm-identified ITs to the right annotation-based category was not always right. For instance, an algorithm-identified IT length might be shorter or longer than the annotation length, and the mapping can only be corrected if the IT identification is correct.

2.4 Optimization

With the goal of maximizing the F_1-score, we optimized parameters for the two algorithms on each dataset. Initial values for centWave on MM14 were selected from the paper Tautenhahn *et al.* (2008); the manual annotations provided a baseline of minimum IT length, height and ppm deviation. Where prior knowledge was absent, liberal parameter grids were explored for parameters like *snthresh* for centWave, or *criticalValue* for Massifquant. Paired parameters, or parameters that were thought to have interactions, were explored simultaneously in two dimensions. For instance, the (min, max) IT length form a natural pair and exhibited interactions in *F*-score performance for centWave. The most important parameters for both algorithms, (*snthresh*, *ppm*) in centWave and (*criticalValue*, *consecMissedLim*) in Massifquant, were searched simultaneously. Their respective *F*-score surface plots exhibited near-concavity, a desirable property for parameter tuning. It appears unique to Massifquant that all *F*-score surface plots had near-concavity. The optimizations were conducted with R (http://www.r-project.org) and Matlab scripts (MATLAB version 7.14.0.739, The Mathworks Inc., Natick, MA). Scripts and detailed procedures to reproduce results are available on GitHub (see Availability and implementation). Other details of the optimization are included in the Supplementary Material. Table 1 compares centWave performance on MM14 based on reported optimized parameters from the original centWave publication and the optimized parameters resulting from this new evaluation. The two different evaluation settings yield similar parameters and F_1-scores, suggesting that this new annotation and evaluation effort are valid. For matchedFilter and MZMine2, all combinations of the suggested ranges for each parameter were exhaustively evaluated (see Supplementary Materials for the full list).

3 RESULTS

3.1 Overall evaluation

As detailed in the Section 2, we developed an independent, open-source implementation of Aberg *et al.*'s TracMass algorithm, and call it 'Massifquant'. The algorithm uses 2D KFs to adaptively find chromatographic ITs in the m/z domain without bucketing the data. We compared Massifquant's ability to sensitively and accurately find ITs with centWave, a sophisticated and well-known algorithm used in the XCMS platform for label-free IT detection, matchedFilter, the original binning-based XCMS method for IT detection and MZMine2, a non-XCMS platform for MS data processing.

We manually annotated ITs in two datasets, chosen to have different characteristics, following a set of rational guidelines. The MM14 dataset is a run of 14 plant metabolites on a lower-resolution UPLC-ESI-QTOF. The MM14 reveals the performance of an IT finder under close-to-ideal circumstances (viz. low sample complexity, good signal-to-noise, good chromatography). The MOUSE sample was run on an Orbitrap mass spectrometer and is typical of many highly complex proteomic analyses. Although chromatographic IT shapes are smooth for high abundance ITs, the intrinsic dynamic range limitations result in greater m/z and intensity variability for lower abundance analytes. The heterogeneity of IT sizes and shapes encountered in the MOUSE data are ideal for discovering the limitations of an IT detection algorithm.

Figure 1 shows that Massifquant reported uniformly higher sensitivity values than centWave, and the *t*-test union of segmented ITs improves Massifquant performance on MOUSE. As for identification reliability, precision was in the same neighborhood for both datasets, yet centWave shows higher sample specificity in MOUSE, as it rarely found a false IT. Massifquant exhibited a better F_1-score on MOUSE, as it identified substantially more ITs than centWave. Both algorithm's MM14 performance is effectively equal for all metrics but sensitivity. The matchedFilter algorithm was only able to identify 33 of the 589 ITs in the MOUSE dataset after optimization over 215 parameter settings. MZMine2's best performance was worse, with only 20 ITs correctly identified under optimal parameter settings (see Supplementary Material). Because matchedFilter and MZMine2 perform so poorly compared with centWave and Massifquant, we omit the results from the charts in this article.

Comparing algorithms' quantitation accuracy is controversial because defining IT boundaries is not clear-cut and in this analysis most error comes from the tails—knowledge afforded because of the evaluation criterion. No statistical test comparing the two algorithms was done, as the spatial components, length, shape, m/z variance, etc. likely create dependence among ITs. Nonetheless, Figure 2 illustrates that Massifquant and centWave quantitation errors are generally in the same small neighborhood.

Table 1. centWave optimization on MM14 improved with identification performance and the parameters are in the same vicinity

Version	ppm	snthresh	peakwidth	peakfilter	F_1-score
Original	30	2	(5,10)	(2, 400)	0.8936
Our evaluation	18.4	2.5	(3,11)	(1, 511)	0.9438

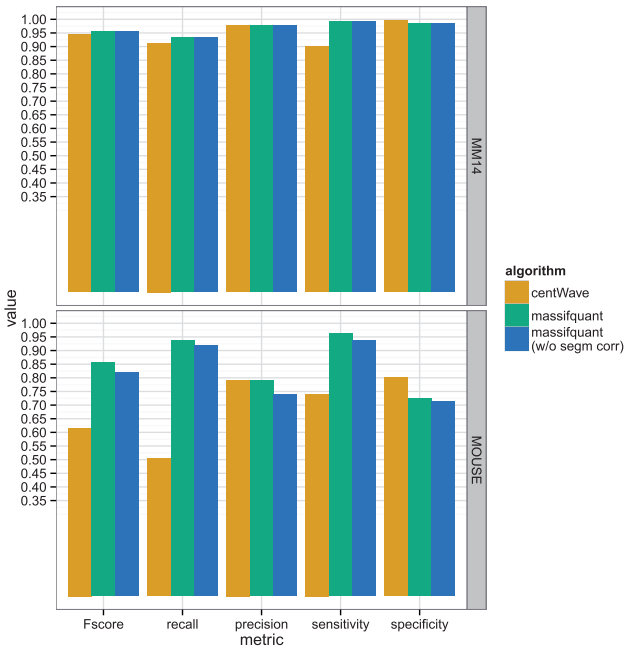

Fig. 1. Optimized performance metrics by dataset and algorithm. This Figure shows the performance of Massifquant without correcting IT segmentation. These figures used reshape2 and ggplot2 R packages (Wickham, 2007, 2009)

3.2 Evaluation by IT type

An evaluation is incomplete without identifying what types of ITs were missed within certain types of samples. For example, both algorithms are perhaps equally excellent at detecting ITs in a simple sample like MM14 with high signal-to-noise (see Supplementary Figure S3). On the other hand, Figure 3 shows that Massifquant excels at finding low-intensity type ITs in the MOUSE complex sample and quantifies them well, whereas these are not identified by centWave.

The 'Evaluation by IT Type' strategy, described in Section 2.3, addresses whether the high number of low-intensity ITs relative to high-intensity ITs in the MOUSE data unfairly benefited Massifquant in aggregate statistics (viz. F_1-score). Figure 4 summarizes the results of IT-typed performance for characteristics thought to vary widely within MOUSE. centWave's IT sensitivity improves as the intensity increases and the estimated ppm error decreases, both in a linear fashion. *Massifquant's* sensitivity varies little across all categories, irrespective of the variable,

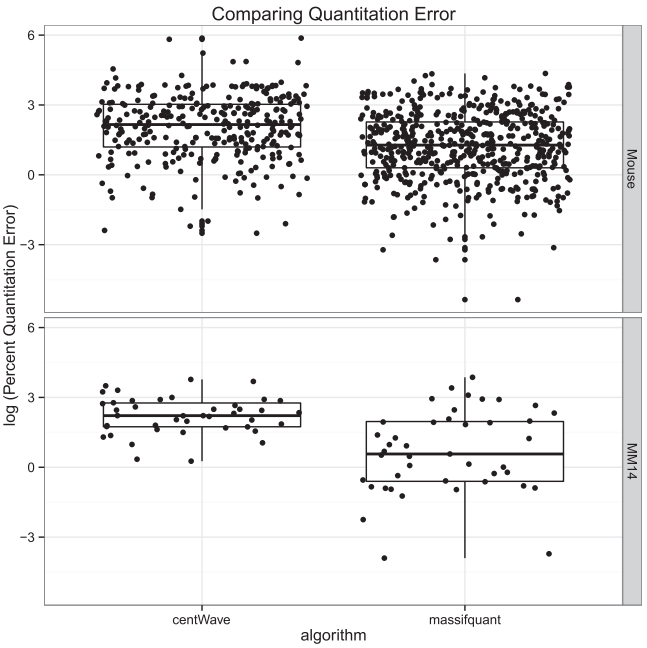

Fig. 2. A comparison of log-transformed percent quantitation errors (log ε) for successfully identified ITs. Massifquant outperforms centWave's quantitation error on both datasets

and without a doubt outperforms centWave. With respect to IT precision, the effect of each variable seems present for both algorithms. Both have similar approximate precision results. Not surprisingly, Massifquant shows improved precision as length, narrowness and max-intensity increase.

4 DISCUSSION AND CONCLUSIONS

In Massifquant, we have implemented an open-source KF-based IT detection algorithm based on Aberg *et al.* (2008). We have evaluated its performance using two manually annotated datasets, and compared the performance of Massifquant with centWave, a wavelet-based IT finder, and matchedFilter and MZMine2, binning-based IT finders. A protocol for how IT detection algorithms should be evaluated has not yet been established; so we first discuss the evaluation process; then, we address algorithmic performance and suitability for use and finally conclude with some thoughts about the use of m/z information in MS IT detection generally.

4.1 The evaluation process

Comparative evaluation of algorithms in MS-omics is often lacking. Smith *et al.* (2013b) and Zhou *et al.* (2012) recently suggested that the quantitative evaluation of IT detection algorithms is long overdue. We believe that the general lack of evaluation is related to the difficulties associated with creating datasets to effectively test these algorithms and also to a lack of clear and explicit metrics for assessing success. To facilitate further efforts in this area, we discuss some of the challenges and successes we met using a manually annotated dataset approach.

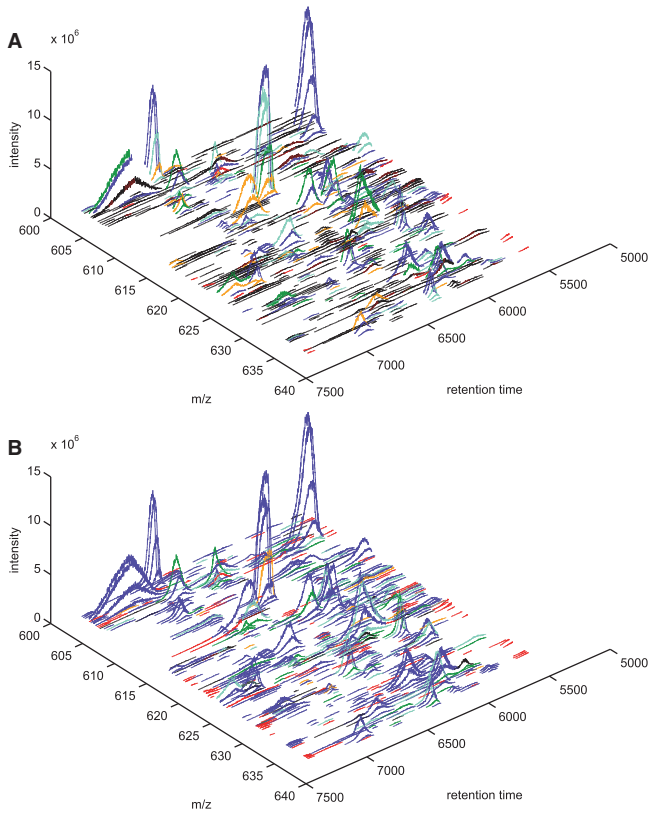

Fig. 3. A comprehensive view of manually annotated ITs on the MOUSE dataset and detected ITs, for (**A**) centwave and (**B**) Massifquant. Correctly identified ITs are color-coded according to the percent quantitation error (ε): dark blue <10%, aqua <20%, green <40%, orange >40%. False ITs are labeled in red; all other noise was excluded. ITs missed by the algorithm (i.e. false-negative results) are labeled black

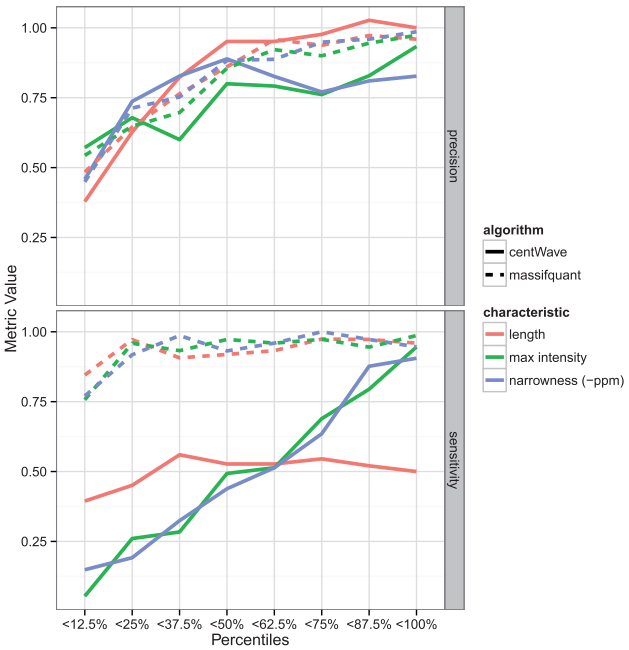

Fig. 4. Isotope trace detection performance across various quantiles for different IT characteristics of the MOUSE dataset. The leftmost percentile bins generally represent the hardest cases for IT detection algorithms (short, low intensity, broad ITs), while bins on the right are generally easier (long, high intensity, narrow ITs). The sensitivity panel is at the IT-level

Hand-annotation, especially of low-abundance ITs, is extremely challenging. It requires concerted effort over a long period. The authors spent several weeks of dedicated effort to annotate the two datasets, and the MOUSE dataset is only a small subset of the complex LC-MS sample from which it was derived. Despite our best efforts to be accurate and consistent, we conclude that the manual annotation process is still somewhat subjective. We simply had to exclude the evaluation of ITs below a certain threshold because we felt human judgment was inadequate for the task. Despite these challenges, the annotation data itself are a useful model for future validation efforts. Moreover, it contains isotopic-level information that could be of use in other projects.

We validated the manual annotation efforts through a holistic visual inspection (see Figure 3 for example) and analysis of histograms of ppm deviation (see Supplementary Figure S4 for example) to ensure that there were no outliers. So, despite the inherent difficulty of manual annotation, we conclude that the endeavor was largely successful. Several aspects of the process are worth considering in more depth: (i) We used semirigid guidelines for annotation that we believe worked well across a variety of ITs with different characteristics. We could have generated and applied strict rules for annotation at the outset, but this may have resulted in even worse systematic bias considering the highly variable ITs we encountered. The proposed guidelines should serve useful for future annotation efforts. (ii) We used a single annotator for both datasets to eliminate person-to-person variability in the interpretation and application of IT criteria. However, tools for community-sourcing annotations would be an interesting solution and has been already been discussed in genomic contexts (Good and Su, 2013). (iii) We used ToppView, the MS viewer associated with OpenMS, to help us find and annotate ITs (Sturm and Kohlbacher, 2009; Kohlbacher *et al.*, 2007). Additional add-ons such as color-coding and flagging of already-annotated ITs and producing a community-based validation would also improve the annotation process.

Among the previous efforts to evaluate IT detection algorithms, we found that most of them focused solely on questions of identification, but lacked in detail of what constituted an 'identified' IT. For IT detection, the identification criterion is critical for fair evaluation, and we additionally argue that the evaluation should probe quantitation accuracy if possible. We evaluated identification at IT and sample levels, and also calculated the percent quantitation error for each IT. The precisely defined metrics may now be more easily used, modified or improved.

This multi-metric evaluation exposes two risks other evaluations take when relying purely on the F_1-score. (i) Precision values show that Massifquant does at least as well if not better at IT identification reliability for MOUSE at low intensity. However, the sample specificity, along with Figure 3, provide

stronger evidence that centWave effectively discriminates low-intensity non-ITs better than Massifquant. Hence, precision and consequently the F_1-score can be misleading. To our knowledge, this is the first evaluation that has proposed a true specificity measure for IT detection, which helps avert wrong conclusions. (ii) By our evaluation standards, and likely others, accurate quantitation does not always imply a favorable IT-detection *F*-score and vice-versa. On the MOUSE dataset, centWave ignores many low-intensity ITs, giving it a low *F*-score; however, the ITs that it does identify are generally quantitatively accurate with a median $\varepsilon = 8.663\%$. Thus, quantitative accuracy is somewhat distinct from IT detection sensitivity or precision.

4.2 Algorithm performance

On the simple dataset MM14, Massifquant showed similar performance to centWave. On a highly complex sample, MOUSE, Massifquant performed much better. In particular, Massifquant excels at finding ITs with a variety of characteristics such as differing intensity, widths and lengths. Massifquant outperforms centWave in IT detection sensitivity across every size and shape of ITs in the complex sample tested. As for reliability, Massifquant is competitive with centWave with the exception that it finds more false low-intensity ITs; the excess false-positive results and multi-modal artifacts are two deficiencies of Massifquant that can complicate downstream analysis in sample-to-sample comparisons. Future extensions of KF IT detection will need to make intensity estimation more robust. An attempt to combine centWave's wavelet intensity estimation with Massifquant has not proven to be effective (see Supplementary section 4). In spite of these deficiencies, both algorithms reported similar quantitation accuracy for the quantified ITs; Massifquant just found far more ITs.

A possible objection to our general comparison is that a large number of small ITs might bias the evaluation in Massifquant's favor. However, Figure 4 removes any suspicion of unfair advantage; even if low-intensity or broad ITs (e.g. first four bins) were removed from the analysis, Massifquant still identifies ITs better on the MOUSE dataset.

As shown in Figure 1, our effort to address the problem of IT segmentation with Massifquant was successful: on the MOUSE and MM14 dataset, the precision increased from 0.7391 to 0.7894 and 0.9185 to 0.9355, respectively. However, some ITs were erroneously combined (Supplementary Figure S2). For algorithmic simplicity, future efforts should attempt to address the IT segmentation problem from within the framework of the KF. Ideally, such an approach would also be more effective than the *ad hoc* method we applied in this study to treat IT segmentation.

4.3 Ease of use

Massifquant parameters can be readily optimized through visual confirmation instead of score-based methods (e.g. *F*-score) that require an annotation. Visual optimization is more time efficient, intuitively simple and almost as accurate. Similar in purpose to Tengstrand *et al*. (2014), the visualization tools at https://github.com/topherconley/optimize-it illustrate precise changes in IT detection induced by differing parameter input. The documentation offers a step-by-step guide on how to optimize Massifquant to new datasets, especially controlling the number of false positives. Further, the score-based method shows a concave *F*-score surface when varying Massifquant's parameters, indicating a predictable parameter behavior (Supplementary Figures S5, S12–S14). Massifquant's appeal is because, at least in part, of the fact that several internal KF parameters are learned from the data—in an initial pre-scan, and then later for each individual IT being tracked.

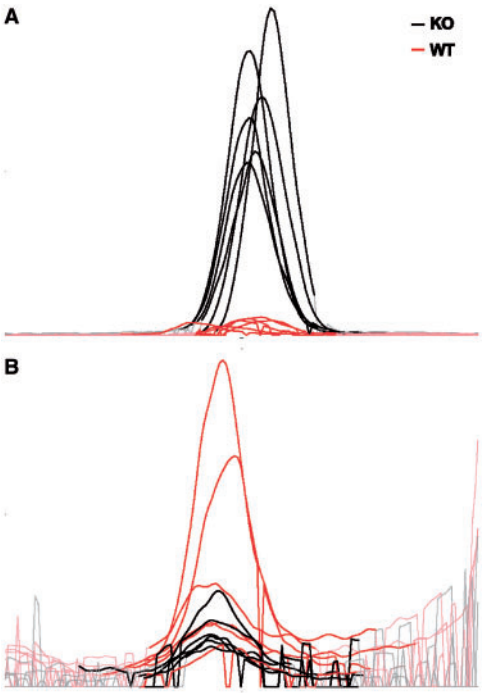

Fig. 5. Massifquant identifies differentially expressed ITs between wild-type versus knockout conditions in the faahKO dataset for (**A**) trivial cases and (**B**) non-trivial cases

Massifquant operates on centroided MS data, which means it can analyze data taken in centroid mode or profile mode (after centroiding), whereas algorithms requiring profile data cannot operate on centroid data because the centroiding process is not readily reversible. Further, running Massifquant is as easy and modular as other XCMS IT detection options. The same differential abundance (DA) workflow applies. Figure 5 illustrates a Massifquant-based DA analysis on the FAAH knock out LC/MS dataset (Saghatelian *et al*., 2004) (see http://bioconductor.org/packages/devel/data/experiment/manuals/faahKO/man/faahKO.pdf for details).

4.4 The use of m/z information in IT detection

Can the success of Massifquant on a complex sample be generalized? ITs in a highly complex sample—particularly low-abundance ITs—are different from ITs derived from a simple mixture: limitations in a mass spectrometer's dynamic range produce much greater intensity variability for ITs from a complex sample. Because of this, at least for mid-to-high mass accuracy/resolution mass spectrometers, m/z measurements will tend to be far more helpful at distinguishing closely eluting species than IT shape. We

found that Massifquant performs at a high level because of its m/z estimation (despite extremely poor intensity estimation). Most IT detection algorithms focus on IT shape, but we suggest that on highly complex samples, an algorithm should be focused mainly on subtle changes in m/z. Algorithms that bin data from closely related ITs to do IT shape analysis lose the richest information available for distinguishing those ITs. Distinguishing convolved isobaric compounds and near-isobaric compounds will, of course, require chromatographic IT shape analysis, but new algorithms will likely see the greatest improvement gains by working to fully use the m/z information found in closely eluting analytes.

ACKNOWLEDGEMENTS

The authors greatly appreciate assistance from Jeffrey Humpherys (BYU, Dept. of Math), William F. Christensen (BYU Dept. of Statistics) and Steffen Neumann (Leibniz Institute of Plant Biochemistry).

Funding: This work was supported by National Science Foundation (0639328), institutional startup funds from Brigham Young University and NSF GRF (DGE-0750759) to R.S.

Conflict of Interest: none declared.

REFERENCES

Aberg,K.M. *et al.* (2008) Feature detection and alignment of hyphenated chromatographicmass spectrometric data: Extraction of pure ion chromatograms using kalman tracking. *J. Chromatogr. A*, **1192**, 139–146.

Bickel,P.J. *et al.* (1975) Sex bias in graduate admissions: data from berkeley. *Science*, **187**, 398–404.

Cappadona,S. *et al.* (2012) Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids*, **43**, 1087–1108.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Good,B.M. and Su,A.I. (2013) Crowdsourcing for bioinformatics. *Bioinformatics*, **29**, 1925–1933.

Kohlbacher,O. *et al.* (2007) TOPP—the OpenMS proteomics pipeline. *Bioinformatics*, **23**, e191–e197.

Michalski,A. *et al.* (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.*J. Proteome Res.*, **10**, 1785–1793.

Pluskal,T. *et al.* (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Saghatelian,A. *et al.* (2004) Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, **43**, 14332–14339.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787. PMID: 16448051.

Smith,R. *et al.* (2013a) Controlling for confounding variables in MS-omics protocol: why modularity matters. *Brief. Bioinform.*

Smith,R. *et al.* (2013b) Novel algorithms and the benefits of comparative validation. *Bioinformatics*, **29**, 1583–1585.

Sturm,M. and Kohlbacher,O. (2009) TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.*, **8**, 3760–3763. PMID: 19425593.

Tautenhahn,R. *et al.* (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, **9**, 504.

Tengstrand,E. *et al.* (2014) Tracmass 2—a modular suite of tools for processing chromatography-full scan mass spectrometry data. *Anal. Chem.*, **86**, 3435–3442.

Welch,G. and Bishop,G. (2006) *An Introduction to the Kalman Filter*. TR 95-041. UNC-Chapel Hill. http://www.cs.unc.edu/~welch/kalman/kalmanIntro.html (7 January 2004, date last accessed).

Wickham,H. (2007) Reshaping data with the reshape package. *J. Stat. Softw.*, **21**, 1–20.

Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Wisniewski,J.R. *et al.* (2009) Universal sample preparation method for proteome analysis. *Nat. Methods*, **6**, 359–362. PMID:19377485.

Yu,T. *et al.* (2009) apLCMS—adaptive processing of high-resolution lc/ms data. *Bioinformatics*, **25**, 1930–1936.

Zhang,J. *et al.* (2009) Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genomics*, **10**, 388–401.

Zhou,B. *et al.* (2012) LC-MS-based metabolomics. *Mol. BioSyst.*, **8**, 470–481.