This item is the archived peer-reviewed author-version of:

Fast and reliable quantitative peptidomics with labelpepmatch

Reference:
Verdonck Rik, De Haes Wouter, Cardoen Dries, Menschaert Gerben, Huhn Thomas, Landuyt Bart, Baggerman Geert, Boonen Kurt, Wenseleers Tom, Schoofs Liliane.- Fast and reliable quantitative peptidomics with labelpepmatch
Journal of proteome research - ISSN 1535-3893 - 15:3(2016), p. 1080-1089
Full text (Publishers DOI): http://dx.doi.org/doi:10.1021/ACS.JPROTEOME.5B00845
To cite this reference: http://hdl.handle.net/10067/1324450151162165141
Fast and reliable quantitative peptidomics with labelpepmatch

Rik Verdonck, Wouter Dehaes, Dries Cardoen, Gerben Menschaert, Thomas Huhn, Bart Landuyt, Geert Baggerman, Kurt Boonen, Tom Wenseleers, and Liliane Schoofs

J. Proteome Res., Just Accepted Manuscript • DOI: 10.1021/acs.jproteome.5b00845 • Publication Date (Web): 01 Feb 2016

Downloaded from http://pubs.acs.org on February 1, 2016

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.
Fast and reliable quantitative peptidomics

with labelpepmatch

Rik Verdonck,* † Wouter De Haes,† Dries Cardoen,† Gerben Menschaert,‡ Thomas Huhn,¶ Bart Landuyt,† Geert Baggerman,§∥ Kurt Boonen,† Tom Wenseleers,† and Liliane Schoofs* †

†Biology Department, KU Leuven, Leuven, Belgium
‡Research Group of Bio-informatics and Computational Genomics, Ghent University, Ghent, Belgium
¶Chemistry Department, University of Konstanz, Germany
§CFP/ceproma, Antwerp University, Antwerp, Belgium
∥Applied Bio & molecular Systems, Vito, Mol, Belgium

E-mail: rik.verdonck@bio.kuleuven.be; liliane.schoofs@bio.kuleuven.be
Phone: +3216330283

Abstract

The use of stable isotope tags in quantitative peptidomics offers many advantages, but the laborious identification of matching sets of labeled peptide peaks is still a major bottleneck. Here we present labelpepmatch, an R-package for fast and straightforward analysis of LC-MS spectra of labeled peptides. This open-source tool offers fast and accurate identification of peak pairs alongside an appropriate framework for statistical inference on quantitative peptidomics data, based on techniques from other -omics disciplines. A relevant case study on the desert locust Schistocerca gregaria proves our pipeline to be a reliable tool for quick but thorough explorative analyses.
Abbreviations

BH: Benjamini-Hochberg
CA: corpora allata
CC: corpora cardiaca
CE: capillary electrophoresis
D0: light isotope of TMAB (0 deuteriums)
D9: heavy isotope of TMAB (9 deuteriums)
DMSO: dimethyl sulphoxide
ESI: electrospray ionization
iCAT: isotope-coded affinity tag
iTRAQ: isobaric tag for relative and absolute quantitation
FA: formic acid
FDR: false discovery rate
FT-MS: Fourier transform ion cyclotron resonance mass spectrometry
LC: liquid chromatography
MALDI: matrix assisted laser desorption ionization
MS: mass spectrometry
TFA: trifluoro acetic acid
TMT: tandem mass tag
TMAB: trimethylammoniumbutyrate
TMAB-NHS: trimethylammoniumbutyrate-N-hydroxysuccinimide
TOF: time-of-flight (mass analyzer)
Q-TOF: quadrupole time-of-flight
SILAC: stable isotope labeling with amino acids in cell culture
TAILS: terminal amine isotopic labeling of substrates
Keywords

isotope tag, label, neuropeptide, peptide, peptidomics, TMAB, PBAN, pyrokinin, hugin, locust

Introduction

Bioactive peptides occur in the whole animal kingdom, from the most ancient cnidarians to the recently evolved mammals.\textsuperscript{1,2} They play key roles as signaling molecules in many physiological processes as well as in various behaviors, for instance, as a peptidergic neurotransmitter or neurohormone, as a toxin, or growth factor.\textsuperscript{3-9}

Therefore, peptides are of considerable biological and medical importance and knowledge of the dynamics of peptides is of critical importance in drug discovery.\textsuperscript{10-12} Most bioactive peptides are synthesized in the cell as large preproproteins from which they are cleaved and modified to generate mature functional entities.\textsuperscript{13-16} Since target cells usually express a variety of receptors, they often react to profiles of circulating peptides rather than to single peptidergic input. In this context, it is often interesting to apply a systems biology approach\textsuperscript{17} and to study the entire peptidome, i.e. the set of all biologically active peptides in a given cell or tissue at a given moment. When studying this entire set of biologically active peptides, mass spectrometry (MS) is the method of choice both for quantification and identification.\textsuperscript{18-20}

The term peptidomics for a new promising “omics” field was not coined until the beginning of this century,\textsuperscript{21-25} when advances in mass spectrometry were made, but gained significant interest when genome projects that delivered comprehensive data pools, took flight.\textsuperscript{26-29} Today, we see the emergence of a multitude of differential peptidomics methods for the quantification of peptides in different physiological or behavioral conditions.\textsuperscript{30} Some of these methods are label free; others apply different kinds of isobaric or stable isotopic tags that are covalently bound to primary and secondary amines provided by the N-terminus and lysine
residues of the peptides. Labels that have been applied for this kind of analysis include succinic acid, formaldehyde and trimethyl-ammonium butyric acid (TMAB). After LC-MS or CE-MS, pairs of labeled peptides can be recognized as co-eluting ions with a mass difference that is an integer multiple of the mass difference between labels. Two peaks that correspond to the light and heavy labeled variant of a peptide, are commonly called a peak pair.

Methods applying stable isotope tags have proven to be particularly robust and reproducible as they allow samples to be compared within one single MS analysis. This paired design requires fewer MS runs, thereby reducing cost without losing statistical power. However, up until now, none of the prevailing mass spectrometry software is equipped with appropriate peak pairing functionalities for non-proteomic peptidomics with a flexible choice of labels. Hence, in the past large parts of the analyses have been done manually. To overcome this limitation, we developed the R package labelpepmatch, which consists of a series of functions that together make up a pipeline from reading in data to a full statistical analysis. It is aimed very specifically at the analysis of endogenously occurring peptides labeled with stable isotope tags, without intending to compete with proteomics software that is typically used for the analysis of labeled enzymatically cleaved (mostly tryptic) fragments of larger proteins (e.g. iCAT).

As a case study, we use the label TMAB to compare neuropeptidomes from solitarious and gregarious desert locusts (Schistocerca gregaria Forskål). Typical for locusts is their population density-dependent swarm-formation, which regularly leads to the destruction of crops in large parts of the world. We are only starting to understand the underlying molecular and physiological mechanisms of their swarming behavior and hypothesize that neuropeptides may play a role. Here, we employ TMAB labeling on the neuropeptidomes obtained from the CC/CA (corpora cardiaca and corpora allata complexes, major neuroendocrine organs) of solitarious and gregarious locusts. TMAB, the label that was used in our study, has been tested extensively by Fricker et al. and has been proven to be an excellent
tool in differential peptidomics.\textsuperscript{31,37–40} The labels featured here do not differ from each other in any way other than a mass difference caused by a different number of incorporated deuteriums. The peptide extract of one experimental condition is labeled with the light ($D_0$) version of the label, the other with the heavy variant ($D_9$).

**Experimental procedures**

**General outline of a typical experiment**

In this study, we discuss the one-by-one quantitative comparison of peptidomes under two different conditions, using stable isotope tags with two different isotopes (TMAB $D_0$ and $D_9$). More complex designs, comparing multiple conditions (like carriage wheel or loop designs) are also possible, but are not discussed here. The general outline of a typical experiment is as follows:

1. **Peptide extraction:** samples (tissues, whole animals) are homogenized and peptides are extracted. There are different suitable extraction methods, including acid methanol extraction and HCl extraction. After extraction, peptide samples are best lyophilized and stored at $-20^\circ C$.

2. **Labeling:** in this second phase, labels are covalently bound to the sampled peptides. After labeling, a light and a heavy labeled sample of different conditions are pooled. Since labels can have batch-specific impurities that can cause side reactions,\textsuperscript{48} it is important that a ”dye swap” be introduced, which implies that samples from one condition are alternatingly labeled with a light or a heavy label.

3. **Purification:** in this very crucial phase, samples are stripped of excessive salts and other impurities. Since labeling protocols usually entail the addition of large quantities of different salts, we advise to either execute an additional washing step, or to apply a
double purification. A number of different reverse phase columns are available for this purpose, but attention has to be paid to their tolerance to the high pH of the samples.

4. Mass spectrometry: samples are subjected to mass spectrometry. Here again a number of possible techniques are available. Since peptidome samples are usually rather complex, the preferred technologies are the so-called hyphenated techniques where mass spectrometry is preceded by another level of sample separation like nano-LC and CE. Our preferred method for samples labeled with TMAB is nano-LC-ESI-Q-TOF, but for samples labeled with formaldehyde or succinic anhydride, CE-MALDI has also been used successfully.\textsuperscript{33–36}

5. Data analysis: hyphenated techniques like nano-LC-MS and CE-MALDI-MS yield complex high dimensional data that can be cumbersome to process manually. In this data analysis step, mass spectra are scanned for peak pairs that correspond to labeled peptide pairs. This is the step where our new pipeline can substitute manual scanning of mass spectra.

The \textit{labelpepmatch} package

There is no shortage of software to analyze mass spectra. A number of very good R packages are available through bioconductor, for example \textit{mzR}\textsuperscript{49} for parsing different MS file formats and \textit{XCMS}\textsuperscript{50} or \textit{MassSpecWavelet}\textsuperscript{51} for finding peaks and matching them across samples.

Furthermore, many software packages, both free and commercial (e.g. Census,\textsuperscript{52} MaxQuant, MsQuant, Mascot Distiller, Progenesis LC-MS, Peaks, Decyder MS, openMS) work very well, and some of them even have functionalities for working with the most common isotopic labels, like TMT, iTRAQ, ICAT or SILAC. However, to the best of our knowledge no package exists that has direct functionalities to detect pairs of stable isotope tagged peptides in an LC-MS or EC-MS space, in a non-proteomic context and regardless of label or which MS method is used.
Our proposed pipeline starts with data that are generated with any among those programs that are available for the interpretation of mass spectra, feature detection and matching peaks across samples. Any list of features, preferentially matched across samples, with m/z, retention time and quantity information can serve as input to our pipeline.

The core functionality of labelpepmatch is to detect pairs of peaks from labeled peptides. The deterministic search algorithm works fast thanks to multi-threading, and detected peaks can be identified using a mass match function that links to databases of known peptides (e.g.\textsuperscript{53,54}). This is especially convenient when no MS\textsuperscript{2} data are present, or when fragmentation or its interpretation is hampered by the presence of labels or modifications (see\textsuperscript{31}). Both for peak pair detection and mass matching, false discovery rates are estimated using mock data and mock databases that are generated based on the real data using restricted randomization procedures. For a detailed description of these procedures, the reader is referred to the labelpepmatch reference manual.

Next to the peak pair search and mass match functions, the package also has built-in inference modules that apply different kinds of mixed models that take run and label effects into account. Since labeled peptide data can be treated much like micro array or RNA-seq data, labelpepmatch is also compatible with the package limma\textsuperscript{55} for transcriptomics. In this manner, the user can take advantage of many additional functions for the normalization, statistical analysis and visualization of the data, as well as the empirical Bayes framework for pooled variance estimation. Apart from a range of functions for data analysis and high level graphics, labelpepmatch contains an inbuilt example dataset (desert locust peptides) and links to peptide databases\textsuperscript{53,54} of a variety of animals. A description of all the objects in the labelpepmatch package along with a detailed workflow is available in the manual and the vignette (see supplementary data). Figure 1 shows an overview of the main functions of the package.
Figure 1: Schematic representation of our analysis pipeline for differential peptidomics data. All functions in the gray zone are contained within the R package. The boxes are objects, and the arrows are functions. Briefly: features are detected with one of the many available dedicated software packages. Next, the resulting list of features is read into an R session. Features are matched to each other with the `pepmatch` function. Here, the resulting peak pairs can be mass matched to a database of known peptides with the `pep.id` function, or identifications from e.g. MS/MS can be added with the `add.id` function. Next, the data can be statistically analyzed for differential expression. Finally, a list of peak pairs can also be passed onto the `limma` package, where a plethora of statistical and graphical functions are available for thorough downstream analysis.
Sampling

Locusts (*Schistocerca gregaria*) were reared at the university of Leuven, Belgium, on a diet of fresh cabbage and rolled oats, under a 14h light, 10h dark regime, at a constant 28 – 30°C. Gregarious locusts were reared at high density (at least 100 animals in a 38cm * 38cm * 38cm cage), while solitarioous locusts were reared in olfactory and visual isolation. All animals had a meal of fresh cabbage in the morning. In the afternoon, CC/CA complexes were dissected from 8 solitarioous and 8 gregarious male locusts, 3 to 5 days post final ecdysis. 200 µl of ice-cold acid methanol solution (90:9:1; methanol:water:acetic acid; v:v:v) was added and samples were sonicated for 5 minutes in a bath sonicator. After centrifugation (9,500 x g for 10 min at 3°C), the supernatant containing the peptides was collected. The resulting pellet was resuspended in 200 µl ice-cold acid methanol solution, sonicated and centrifuged again. The supernatant of both extractions was pooled, and lyophilized until all methanol had evaporated. The remaining aqueous solution was de-lipidated using an equal volume of n-hexane and further lyophilized until dry.

Labeling

TMAB labels were produced at the university of Konstanz according to prescribed protocols. Samples were labeled with two stable isotopic variants: the light (D₀) and heavy (D₉) species of TMAB differ by 9 neutrons, located in the hydrogen atoms of the methyl groups. The labeling procedure starts with the suspension of the dry peptide extracts in 100 µl of 5% CH₃CN (acetonitrile) and 0.1% TFA in ultra-pure water (Milli-Q), followed by sonication and centrifugation. As the optimal labeling reaction takes place at a pH between 9 and 9.5, the solution is buffered with 200µl of aqueous disodium hydrogen phosphate (1.5 M Na₂HPO₄, pH 9.5). The labels, either D₀ or D₉, are added in the TMAB-NHS form as a 1.5 M TMAB-NHS solution in water-free DMSO. 4µl label solution is added to the peptide extracts and is incubated for at least 10 min. at room temperature. Afterwards, pH is measured with a glass electrode pH probe and if necessary adjusted to 9 - 9.5 by adding...
Figure 2: Schematic representation of experimental procedures. Each condition, in this case the brain of a gregarious and a solitary locust, is labeled with a different label. Next, samples are pooled and analyzed on the mass spectrometer. In a next step, peak pairs have to be detected in this spectrum.
small amounts (1 - 2 µl) of 1 M NaOH. These procedures (adding label and pH check) are repeated an additional six times with 15 minutes intervals. After 7 cycles of adding TMAB-NHS and pH adjustments, the mixture incubates for another 2 hours at room temperature. Next, the reaction is ended by quenching the excess of label with 30 µl of 2.5 M aqueous glycine to each sample, incubating 40 min. at room temperature. After readjustment of pH to 9.5 with 1 M NaOH, 3 µl of 2 M hydroxylamine (NH₂OH) in water-free DMSO is added in order to remove labels bound to threonine, serine and tyrosine. This procedure is repeated after 10 and 20 minutes. At this stage samples can be pooled and stored at −80°C. Our experimental design was a simple two condition comparison with 8 biological replicates. Four samples from solitarious locusts were labeled with D₀ and paired with D₉ labeled gregarious samples, and four other samples the other way around. After labeling, the samples were desalted using OasisTM HLB cartridges (Waters), vacuum-dried and stored at −20°C.

**Nano-LC-ESI-Q-TOF mass spectrometry**

Samples were dissolved in mobile phase A (2% CH₃CN and 0.1% FA) and analyzed by means of Nano-LC-ESI-Q-TOF. Samples were first separated on nano-LC (Dionex UltiMate 3000 Dual™ LC System; 75 µm x 15 cm Dionex Nano Series™ C18 column) with a 250 nL per minute flow rate. The applied linear gradient started from 5% mobile phase B (90% CH₃CN in water) in mobile phase A to 50% mobile phase B in 45 min at a flow rate of 250 nL/min. The column was washed with a steep gradient towards 100% mobile phase B between every two runs. Mass spectrometry was executed on microTOF-Q (Bruker Daltonics) in MS mode.

**Data analysis with Progenesis LC-MS and labelpepmatch**

**Progenesis LC-MS**

Bruker .d folders were imported in Progenesis LC-MS with the appropriate plugin. Run H was automatically selected as a reference, and all runs were aligned against it. No vectors
were manually added. Peak picking was done with the automatic method, which uses a noise estimation algorithm to the determine the noise level in the data [sic]. Sensitivity level was set to the default value of 3, with a minimum peak width of 1 min. All runs were selected for peak picking. Normalization was disabled and a simple two condition setup was chosen. Note that this setup is of no influence on our analysis, since we do not compare peak intensities between mass spectra. It is however a step that cannot be skipped in Progenesis LC-MS. The high number of 36,000 automatically detected features was brought down to 3425 by excluding the more than 20,000 single charged features, features with more than 8 charges, and features with deconvoluted masses higher than 4 kD. Finally, because the human eye is often able to quickly detect problematic feature detections, the remaining features were manually curated by the following criteria:

- Features that were encountered in less than 4 out of 8 spectra were removed. This allows for the selection of on-off signals where we would expect a feature to pop up in every other run, while we discard a lot of uninteresting features that only show up in a minority of the runs.

- Double hits, where a single feature is detected multiple times over the course of its elution, were merged into single features.

- In some cases, the automated feature detection algorithm in Progenesis was not able to correctly recover the charge, or the first isotope peak. This was especially true for peaks with a high number of charges, or when heavy background noise was present. These were corrected where possible, or else omitted.

- Some feature detections were edited to better include the entire area of the peak over all runs, or to exclude signal caused by neighboring peaks.

- Some very obvious single charged peaks were re-added to the peak list.
In the end, a final set of 1425 candidate peaks was exported as a .csv file. All these procedures can be found in a series of screenshots in supplementary figures.

**Labelpepmatch**

An R session was started and the *labelpepmatch* package was loaded. The .csv file generated by Progenesis LC-MS was read in using the *read.progenesis* function. Peak pairs were detected with *pepmatch* and mass matched to a database of known and predicted *Schistocerca* peptides using *pep.id*, with a set of detection and identification parameters based on the specifications of the MS hardware. Differential expression was inferred with a simple linear model with label as a fixed effect (model 0), using *lpm_linearmodel*. A detailed overview of this workflow, along with many extra interpretations can be found in the vignette of the package (see supplementary data).

**Further downstream analysis with *limma***

Bioconductor *limma* is a package that provides tools for the analysis of micro array data, and since labeled peptide data are structurally very comparable to microarray data, it is an excellent extension to our pipeline. To this end, we provide a direct link with the *limma* environment through the *lpm_make.RGList* function, which creates an object of the *limma* class *RGList*, feeding directly into most *limma* functions. This creates the opportunity to take the analysis further than the linear model offered by *labelpepmatch*, and use different kinds of normalization, Bayesian estimation of variance, and a plethora of extra plotting functions. We used the *eBayes* function to calculate moderated t-statistics for linear model fits on normalized data. We compared 3 models: raw unnormalized data (model 1), within run median normalized data (*normalizeWithinArrays* function) (model 2), and within run median normalized, and between run quantile normalized data (*normalizeBetweenArrays* function) (model 3). All these models were compared to the linear model described in the previous section (model 0). (See supplementary table 2).
Results and discussion

Peak pair detection and peptide identification

There are two main explanations for the staggering initial number of 36,000 automatically detected features. First, the mass spectra suffered from heavy streaking with regular 1 Da intervals, leading to the detection of 20,000 single charged peaks, and secondly, with non-stringent feature detection parameters, Progenesis has the tendency to detect the same feature multiple times, with slight differences in box size.

After successful feature detection, the rest of the data analysis only takes a couple of minutes on a 8-core pc. This is a significant improvement compared to earlier methods that included a lot more manual analysis. However, the biggest advantage of the labelpepmatch package is to have the entire data analysis contained in a pipeline with regular intermediate quality checks and graphics (as illustrated in the package vignette). This allows for a deep understanding of the structure of the data.

In the end, a total of 118 peak pairs were found in all eight runs without quantity threshold. The initial input data contained 1425 features, which begs for the explanation of about 1000 non-labeled features. Because in earlier studies, peak pair detection was not automated,31–40 there is little information available about which proportion of total features is usually part of a peak pair. Most of the non-peak pair features only occur under one ionization state, typically with two, three or four charges. They also have an average lower intensity than the peak pairs (both median and mean more than twofold), but the profile of their masses and retention times is the same (see supplementary figures 1-3). Perhaps part of them are missed peak pairs (e.g. with too low intensity), but others may also be non-peptide metabolites, unlabeled precursors, signal peptides and proteolytic products, or solvent and column contaminants.

Among the non-peak pair features, ten could be mass matched to known Schistocerca peptides (5 ppm threshold), but this does not exceed the estimated false discovery rate,
which was very high due to the high number of features. Hence, a good proportion of these identifications are expected to be false positives. However, it is interesting to remark that the pyroglutamated peptides corazonin, adipokinetic hormone II and hypertrehalosaemic hormone are among them. These 3 peptides have been observed in the locust CC before, and are expected not to interact with the labels due to shielding of the N-terminal and the absence of lysine residues. Peptides that are identified as peak pairs usually do not show up in the unlabeled fraction, indicating that labeling was complete, as described in the literature. The detection of a somewhat limited number of peak pairs out of a large set should therefore be considered an advantage: the pepmatch function is able to quickly extract signal out of a noisy background.

A total of 84 peak pairs with a minimal quantity of 256 in at least one feature were found in all eight runs. Peak pair detection FDR was estimated around 2%. Of these peak pairs, 15 could be mass matched (mass difference < 5 ppm) to 11 known Schistocerca peptides using the pep.id function. IRP-co-peptide, AKH-JP-1 and AKH-JP-2 were found both as single and double charged cations. PK-5 was found both as a double and a triple charged cation. False discovery rate estimates for identification by mass match ranged from 6% to 16%.

It is not uncommon in peptidomics studies to find a high number of initial features often with high total ion counts, only to then identify a limited subset of them. Here again, the identity of the 69 peak pairs that did not get a mass match remains elusive. Some of them may be (neuro-) endocrine peptides that have yet to be discovered, others may be small precursors, or known peptides with modifications, and yet others may be of intracellular origin. Many peptidomic studies have detected a large number of peptides from cytosolic, nuclear and mitochondrial origin, but given the absence of a Schistocerca gregaria reference genome, it is very difficult to predict which non-endocrine peptides to expect. A list of expected and observed neuropeptide identifications can be found in supplementary table 1.

In this study, identifications solely relied on mass match, but in many cases MS/MS can
provide more convincing identifications. However, extra care should be taken when applying MS/MS on labeled peptides. For example, TMAB labeled peptides require higher collision energy, and the straightforward identification is difficult because of the possible fragments of the label. Some search engines, like Mascot (Matrix Science) allow some labels as fixed modification (e.g. GIST-Quat for TMAB), but given the increased chance of false positives when many possible modifications are used, all MS/MS identifications in TMAB studies are manually curated (Wardman 2010, Gelman 2010).

In the case when additional identification methods, like MS/MS, are used alongside mass-match, it is perfectly possible to manually add identifications to the 'pepmatched' object using the add.id function, and proceed with the analysis as described in the paper and the vignette.

**Detecting differentially expressed peptides**

Throughout this pipeline, we worked with raw intensities (height of the peaks). In our study, the heavy label channel had an overall lower intensity, which was consistent over all runs. However, using the right statistics (lpm_linearmodel), or normalizations (limma package), we could easily pick up the biologically relevant signal. Since this label specific effect has not been mentioned in earlier studies, we assume it is specific to our experiment, or to the batches of labels that we used. The fold changes compare well over different runs, and it is consolidating that differently charged ions from the same peptide show very congruent fold changes within one run, which is very apparent looking at the heat maps.

Using a general linear model with label as a fixed effect (model 0, generated with lpm_linearmodel function), 15 differential features were found (p < 0.05 after BH FDR), 10 of which were more abundant in solitary than in gregarious animals. Some of these differential features represent the same peptides with different numbers of charges. Most noteworthy changes are a group of pyrokinins (PK-4 and PK-5) that seem to be more abundant in solitarius animals, and a co-peptide of ion transport peptide (ITG-co peptide) that seems to
Figure 3: Heat map on the contrasts of the residuals of the linear model. Green means higher in solitarious locusts. Notice how pyrokinin 5 strongly clusters together with other non-identified peptides. Because of possible label bias, standard heat maps for unnormalized data would show patterns influenced by labeling design. However, our linear model takes label as a fixed effect. The resulting residuals are the net effect of the treatments, and are hence a good statistic to plot in our heat maps. The lpm_heatmap function either plots these residual effects separately, or their contrasts between two conditions. The first option can be helpful to look for outlying samples, while the second option is best suited for the visualization of correlational patterns in the expression of different peptides. The volcano plot is another common visualization for differential expression. The blue zone is for BH corrected p-values below 0.05, the green zone below 0.01. Red dots are mass matched peptides. This figure is generated with the lpm_volcanoplot function. Clicking on the dots prints the identity of the peak pairs in the R terminal.
be higher in gregarious animals.

Proceeding in limma, three linear models were compared. These are all based on moderated t-statistics through parametric empirical Bayesian inference. A simple linear model fit that did not take into account the label effect (model 1) was not able to detect any significant differences after BH multiple testing correction. This is not surprising given the rather strong label bias. Model 2, with within run median normalization, yielded only two BH adjusted p-values under 0.05, and another 6 under 0.1. Finally, model 3, with the very drastic quantile normalization, yielded no BH adjusted p-values under 0.05, but 10 under 0.1. Fold change estimates were very congruent between all models (supplementary figure 4), but unadjusted p-values show less clear correlational patterns (supplementary figure 5).

Supplementary table 2 displays the list of all peak pairs that passed the most stringent peak pair detection criteria, with all log fold changes and (adjusted) p-values for all models.

It is interesting to notice that a simple linear model that takes label effect into account per feature, turns out to be more powerful than the limma approach with normalizations based on the entire channels. This means that, at least in our experiment, label effect is best estimated per feature. In any case, we want to advocate for looking at more than p-values, especially when balancing on the edge of statistical power in the light of multiple testing. Expression profiles of peptides that originate from the same precursors are by definition non-independent. As a result, classical multiple testing correction procedures (even BH) will be too conservative, and they will fail to detect real differences with adequate inference power. The fact that the limma pipeline yields borderline p-values is actually not of big concern, since many of them are repeats of the same peptide, blowing up the p-values just by the nature of BH procedures. The fact that the peptides are from the same family and they show similar trends in the same individuals, with strong correlations in their fold-changes across samples, is ample extra support for our inference about differential expression. Hence we argue that in many cases, both model 0 and model 2 will be appropriate inference methods, and we encourage the use of both methods alongside each other, combined with a close look
into the correlational structure of the data. The *labelpepmatch* package offers a suite of functions that provide exactly this information.

**Differentially expressed peptides in the desert locust neurohaemal complex**

Not much is known about differential expression of neuropeptides between the two phases of the desert locust. In another species of swarming locust, *Locusta migratoria* HPLC was used to show that brain levels of adipokinetic hormone (AKH) were different between two phases.61 In another study in *Schistocerca gregaria*, two neuroparsins were shown to be differentially expressed in the fat body and the brain.62 A transcriptomic study confirmed these differences in the expression of precursors of two neuroparsins and AKH in entire brain samples.46 Apart from neuroparsins and AKH, there is one enigmatic 6 kD peptide known as phase related peptide, that is found in increased concentrations in the gonads and hemolymph of gregarious desert locusts, but it shows no direct similarities to any known peptide, nor is its function known.63,64

**Pyrokinins or -FXPRL-amides**

Pyrokinins are neuropeptides that are characterized by an FXPRL-amide C-terminal. They were first isolated from head extracts of the cockroach *Leucophaea maderae*,65 and turned out to have myostimulatory effects on the hindgut (a *kinin* like action, hence their name). Around the same time, they were also discovered to have stimulatory effects on pheromone biosynthesis in the lepidopteran *Heliothis zea*,66 where they were called pheromone biosynthesis activating neuropeptide (PBAN). Later, these peptides have also been demonstrated to be involved in cuticular tanning,67 diapauze induction,68 pupariation69 and feeding.70

Like many neuropeptides, pyrokinins are typically part of larger precursors that encode multiple neuropeptides. In *Drosophila melanogaster*, these are precursors encoded by the *CAP2b/MT* gene that contains two periviscerokinins and a pyrokinin (Drm-PK1),71 and the
hugin gene that contains an ecdysis triggering hormone like peptide and a pyrokinin (Drm-PK2). Transcripts from the hugin gene are speculated to be orthologous to mammalian neuromedin (NmU25 and its shorter form NmU8),\textsuperscript{72} and are involved in the regulation of feeding behavior.\textsuperscript{70}

In heterometabolous insects, multiple different pyrokinins are often found, and it is generally assumed that they are orthologous to the Drosophila hugin gene and the lepidopteran pban gene. Also in locusts, many different pyrokinins have previously been observed in the CC and CA.\textsuperscript{57,73} Since the sequences EGDFTPRLa (PK-4) and DSSEDWAQPFPVRLa (PK-5) are not known from any transcripts or EST’s, it is for now unknown if they originate from the same precursor. However, their obvious clustering together in the heat map (with Pearson r > 0.9) suggests that they are either co-encoded, or very tightly co-regulated. This was already proposed by Clynen et al. based on tissue distributions.\textsuperscript{57,73} This finding also leads us to think that most of the unidentified features that cluster strongly together with these two identified pyrokinins are most likely larger precursors of the these peptides.

Some of the putative functions of pyrokinins correspond well with the differences between the two extreme phases of Schistocerca gregaria. For example, in Drosophila melanogaster hugin-expressing neurons appear to comprise a neural circuitry in the brain that regulates taste-mediated feeding behavior.\textsuperscript{70,74} In locusts, the response to gustatory and nutrient signals is indeed very different between the two phase, where the less active solitarious phase is a very choosy feeder while gregarious locusts tend to feed on a variety of toxic or unpalatable plants for predator avoidance.\textsuperscript{75–77}

Another interesting parallel can be drawn with the white spotted tussock moth Orgyia thyellina. This lepidopteran exhibits seasonal polyphenisms in morphological, physiological, behavioral and reproductive traits. These polyphenisms seem to be mediated through FXPRl-amide neuropeptides in neurosecretory cells of the subesophageal ganglion whose axonal projections reach the CC.\textsuperscript{78} Locust phase polyphenism is not seasonal, but the two phases also differ in morphology, physiology, behavior and reproductive strategy.
ITP-co-peptide

ITP-co-peptide is a 9 amino acid peptide that is co-encoded with ITP, which is a larger peptide (about 8 kD) that cannot be observed with the experimental setup used in this study. At this moment, no function is known for ITP-co-peptide. In locusts, it regulates ion and fluid transport over the ileum, stimulating water and ion re-absorption while inhibiting acid secretion. In this way it plays an important role in the homeostasis of hemolymph volume and composition. In Drosophila melanogaster ITP was recently shown to be of importance in circadian rhythm, mediating the timing of locomotor activity along the day. Knocking down ITP in clock cells increased nocturnal activity.

The upregulation of ITP-co-peptide in gregarious locusts is interesting, but difficult to interpret. First of all, the support for its identification is only a single mass match, and secondly only the co-peptide and not ITP itself was observed. It is however interesting to remark that gregarious locusts are diurnal, while solitary locusts are more likely to make crepuscular or nocturnal flights. Gregarious locusts should also be more adapted to retain fluids to prepare for longer diurnal migratory flights. Hence, a potential difference in expression of a putative mediator of circadian rhythms and electrolyte homeostasis is definitely worth a closer look in future studies.

As a final remark, it cannot be ruled out that the observed differences in expression of both pyrokinins and ITP-co peptide are not a cause, but rather a consequence of the phase state of these animals. Locusts from both phases are being reared under different conditions, and although we have done efforts to keep as many environmental variables as equal as possible, it is difficult to assess to which extent possible local differences in air humidity, food availability and stress contribute to their peptidomic profiles. Nevertheless, we suggest that these interesting parallels with other insects, both for pyrokinins and ITP, are worth further investigations through functional experiments with e.g. RNAi knockdown.
Conclusion

This study is the first one applying TMAB to invertebrate peptides, and we hereby confirm that TMAB is a suitable label for the analysis of entire peptidomes, even when peptide concentrations are low.

The *labelpepmatch* package and the here proposed data analysis pipeline drastically improves data analysis of stable isotope tag peptidomics. Thanks to the many visualization functions and quality checks, the data can be studied in great detail, and a clear signal can be obtained, even if mass spectra are noisy, or label biases are present. We demonstrate that a simple linear model can be very powerful in detecting differentially expressed peptides, but also tested techniques borrowed from other -omics disciplines can be successfully incorporated.

Finally, we demonstrate that crowd-reared desert locusts have altered neuropeptide profiles in their neurohaemal organs compared to their conspecifics reared under isolation conditions.

Acknowledgement

The authors thank the staff of the KULeuven Facility for Systems Biology based Mass Spectrometry (Sybioma) and the University of Antwerp Centre for Proteome Analysis & Mass Spectrometry (Ceproma) for assistance with mass spectrometry. Thanks to Lloyd Fricker and colleagues for our training in TMAB labeling and valuable discussions. Thanks to Maarten Michiels who provided insight and expertise in programming work, and Uli Ernst and Katja Nowick for valuable comments on the manuscript. Also many thanks to the R community for being a resource of help and support in the development of this package. This project was funded by BOF (GOA/11/002) and FWO Flanders (G.0437.10 and G.0588.11N). RV was funded by FWO Flanders during this work. WDH is funded by the KULeuven research fund.
Supporting Information Available

- vignette

This vignette displays the entire workflow of the analysis of locust peptide data from reading in data to full statistical analysis. It is an integral part of the package, and can be found here:


- reference manual

The manual is a concatenation of all helpfiles. It contains detailed information about how different functions of the package work, and interact with each others output. It can be found here:


- supplementary figure 1

Histograms of the deconvoluted molecular weights of all detected peaks.

- supplementary figure 2

Boxplot of the log2 transformed quantities of peaks within peak pairs, and non-peak pairs for all 8 runs.

- supplementary figure 3

Two dimensional representation of all peaks, with their retention times plotted onto their deconvoluted masses.

Comparison of log fold changes calculated from all compared models.

- supplementary figure 5

Comparison of the logarithm of uncorrected p-values calculated from all compared models.

- supplementary figures 6 - 16
Screenshots from the pre-processing stage of the data analysis using the commercial software package Progenesis LC-MS.

- supplementary table 1
  All peptides that have previously been observed in *Schistocerca gregaria corpora cardiaca* (CC), *corpora allata* (CA), and mass matched in our study within peak pairs (PP) or outside of a peak pair (N-PP).

- supplementary table 2
  All peak pairs found in all 8 runs with their identifications through mass match, their deconvoluted molecular weight and their number of labels, alongside their possible differential expression according to the four compared models.

- supplementary method: workflow
  This is an R script that contains the entire workflow of the data analysis used in the paper, with a couple of extra lines of code that generated all supplementary figures. It also starts with some details about how to install the package from github.

This material is available free of charge via the Internet at http://pubs.acs.org/.

References


based transcriptomic analysis of differences between long-term gregarious and solitari-

Verdonck, R.; Vanden Broeck, J. Critical role for protein kinase A in the acquisition
of gregarious behavior in the desert locust. *Proceedings of the National Academy of

(48) Fricker, L. D. Limitations of Mass Spectrometry-Based Peptidomic Approaches. *Journal

(49) Martens, L. et al. mzML—a community standard for mass spectrometry data. *Molecular
& cellular proteomics : MCP* 2011, 10, R110.000133.

(50) Smith, C.; Elizabeth, J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing
mass spectrometry data for metabolite profiling using Nonlinear Peak Align-

(51) Du, P.; Kibbe, W. a.; Lin, S. M. Improved peak detection in mass spectrum by incor-
porating continuous wavelet transform-based pattern matching. *Bioinformatics* 2006,
22, 2059–2065.

(52) Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R. A quantitative analysis software tool

(53) Fälth, M.; Sköld, K.; Norrman, M.; Svensson, M.; Fenyö, D.; Andren, P. E. SwePep, a
database designed for endogenous peptides and mass spectrometry. *Molecular & cellular

(54) Liu, F.; Baggerman, G.; Schoofs, L.; Wets, G. The construction of a bioactive peptide


(69) Xu, W. H.; Denlinger, D. L. Molecular characterization of prothoracicotropic hormone and diapause hormone in Heliothis virescens during diapause, and a new role for diapause hormone. *Insect Molecular Biology* 2003, 12, 509–516.


