Bioinformatics, 31(13), 2015, 2228–2231
doi: 10.1093/bioinformatics/btv105
Advance Access Publication Date: 23 February 2015
Applications Note

Systems biology

specL—an R/Bioconductor package to prepare peptide spectrum matches for use in targeted proteomics

Christian Panse*,†, Christian Trachsel†, Jonas Grossmann*,† and Ralph Schlapbach

Functional Genomics Center Zurich, Winterthurerstr. 190, CH-8057 Zurich, Switzerland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on October 15, 2014; revised on February 11, 2015; accepted on February 12, 2015

Abstract

Motivation: Targeted data extraction methods are attractive ways to obtain quantitative peptide information from a proteomics experiment. Sequential Window Acquisition of all Theoretical Spectra (SWATH) and Data Independent Acquisition (DIA) methods increase reproducibility of acquired data because the classical precursor selection is omitted and all present precursors are fragmented. However, especially for targeted data extraction, MS coordinates (retention time information precursor and fragment masses) are required for the particular entities (peptide ions). These coordinates are usually generated in a so-called discovery experiment earlier on in the project if not available in public spectral library repositories. The quality of the assay panel is crucial to ensure appropriate downstream analysis. For that, a method is needed to create spectral libraries and to export customizable assay panels.

Results: Here, we present a versatile set of functions to generate assay panels from spectral libraries for use in targeted data extraction methods (SWATH/DIA) in the area of proteomics.

Availability and implementation: specL is implemented in the R language and available under an open-source license (GPL-3) in Bioconductor since BioC 3.0 (R-3.1) http://www.bioconductor.org (Trachsel *et al.*, 2015). A vignette with a complete tutorial describing data import/export and analysis is included in the package and can also be found as supplement material of this article.

Contact: cp@fgcz.ethz.ch or jg@fgcz.ethz.ch

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

1 Introduction

Targeted proteomics is a fast evolving field in proteomics and was elected as method of the year in 2012 by the *Nature* journal (Editorial, 2013). Methods such as Selected Reaction Monitoring (SRM), Sequential Window Acquisition of all Theoretical Spectra (SWATH) MS (Gillet *et al.*, 2012) and DIA represent promising perspectives for the identification and quantification of peptides and proteins. In these targeted proteomics methods a predefined set of fragment ion signals (transitions) are extracted over the chromatographic time. All methods have in common: the need for precise MS coordinates composed of precursor mass, fragment masses and retention time. These coordinates are stored in spectral libraries. Spectral libraries are generated from shotgun MS/MS data. A protocol on how to build high-quality assay libraries can be found in Schubert *et al.* (2015). Starting from one spectral library, different assays can be extracted according to the requirements of different types of mass spectrometers. Here, an R/Bioconductor package is presented, which creates specific assay panels from peptide identification results, e.g. from BiblioSpec files (Frewen and MacCoss, 2007) or Mascot result files. specL is a versatile set of functions that can easily be integrated into existing commercial or open-source software pipelines for targeted proteomics data analysis. Examples of currently available pipelines are ProteinPilot combined with PeakView (Lambert *et al.*, 2013), Spectronaut (Bernhardt *et al.*, 2012) or OpenSWATH (Rost *et al.*, 2014).

2 A typical targeted proteomics pipeline

The usual starting point of a targeted proteomics experiment is an MS/MS spectral library of identified peptides that should be quantified. MS/MS spectral libraries can be downloaded from publicly available databases (Picotti *et al.*, 2013; Schubert *et al.*, 2013). However, because not all organisms or treatments are available, the creation of libraries from the scratch is necessary. The peptide identification process of database search engines usually results in redundancy. In protein identification this is normally less problematic, while for assay panel building for targeted proteomics experiments, the search results must be redundancy filtered and for each peptide ion the most representative peptide-spectrum-match (PSM) has to be selected. To create this non-redundant input file, we use the BiblioSpec algorithm of the Skyline software suite (MacLean *et al.*, 2010) to filter peptide identification results. This 'Skyline workflow step' outputs two SQLite files named `'*.blib'` and `'*.redundant.blib'`. These files can be used as input for our specL package. During the read bibliospec step, the SQLite file is converted into a psmSet object. The flexible filter options of specL can then be employed to create a customized assay for all peptide identifications in the spectral library, tailored to the characteristics of different mass spectrometry instruments. A graphical representation of this workflow and its individual steps is shown in Fig. 1. The specL output file functions as straight forward input (assay library), e.g. for the BiognoSYS Spectronaut software or, with minimal reformatting, also for PeakView of AB Sciex. Alternatively, it can be used for a script-based construction of SRM/MRM assays.

3 Functionality

3.1 Read BiblioSpec files

The function `read.bibliospec` performs an SQL query on the SQLite files generated by BiblioSpec using the `RSQLite` package. This function is required to import Skyline spectral library files into R. This functionality is tailored for redundancy-filtered input files. It is also possible to directly load Mascot result files using the CRAN `protViz` package (Panse *et al.*, 2015) import functionality as described in the vignette of `specL` in more detail.

3.2 Protein annotation

The protein–PSM mapping information is, by default, not stored by BiblioSpec. Therefore, `specL` provides the `annotate.protein_id` function that uses the R specific grep command to 'reassign' the protein information. It is important to have the flexibility of a regular expression to accommodate a digest pattern. To associate the proteins, a corresponding `fasta` object has to be loaded into R using, e.g. `read.fasta` from the `seqinr` CRAN package. It does not have to be the same FASTA file as the one used in the original database search. The `specL::annotate.protein_id` function may require intensive computing for large FASTA formatted files or potentially big psmSet objects and is therefore designed to utilize a multi core architecture by using the `BiocParallel` package (Morgan *et al.*, 2015).

3.3 Generation of the ion library

The function `genSwathIonLib` generates a specL object as container for the peptides of the original spectral library (psmSet) based on the applied parameters. It is usually recommended to use the most intense and the same number of transitions for each peptide in the spectral library. To carry out this selection we use the parameters *fragmentIonRange* and *topN*. The first parameter defines a range with a minimal and a maximal number of matched fragment ions per precursor and the parameter *topN* selects the *n* most intense fragment ions for each assay. A third important parameter, the fragmentation function, allows to specify the type and charge of ions to be used in the assay. A meaningful default set of parameters is described in more detail in the package vignette.

3.3.1 Normalizing retention times using iRT

retention time is an important factor in targeted data extraction. However, retention times are not easy to transfer between different chromatographic systems. To make a transfer possible and to account for inter-run shifts in retention time, BiognoSYS proposes the iRT normalization (Escher *et al.*, 2012) based on specific iRT peptides. A set of control peptides (good ionization properties, good fragmentation characteristics, completely artificial) that spread over the whole retention time range of the LC gradient are spiked into each experiment. These peptides can later be used to apply a linear regression model to transform all measured retention times into an independent retention time scale. If the identification results contain iRT peptides specL, supports the conversion to iRT scale. For this purpose (if the identification results are based on multiple input files) it requires the redundant BiblioSpec file including all iRT peptides from all measurements. The regression model is calculated for each measurement individually. The original filename is automatically identified for the most representative spectrum in the R-object so that the respective linear model can be applied to normalize the retention time. The *data.fit* parameter handles the individual normalization, if the redundant BiblioSpec object is provided. In case no iRT peptides were spiked or not enough iRT peptides were identified, the retention time regression is not applied and specL will use measured retention time. It is also possible to use a different set of retention time peptides by changing or extending the iRT peptide table (provided by the package).

3.4 Export ion library

The generated ion library is entirely designed as an R language S4 class and can therefore be exported to an ASCII file by using the generic `specL::write.spectronaut` method. Afterwards the file can be used for targeted data extraction in other software packages, e.g. Spectronaut. Additional export functions will be available in future releases.

3.5 Additional features

As additional features, to review or summarize the input of a spectral library (psmSet object) or the assay library (specL object), the functions `specL::plot` and `specL::summary` have been implemented. The `specL::summary` function provides information about the number of precursors in a psmSet, the individual contributions from each raw data file. For a specL object, the summary provides a description of the applied parameters, the number of found iRT peptides and the frequency of the number of transitions per precursor. This allows to quickly review the specL library with respect to the applied parameters and iRT normalization. The `specL::plot` function applied to a psmSet object provides an LCMS map which is a

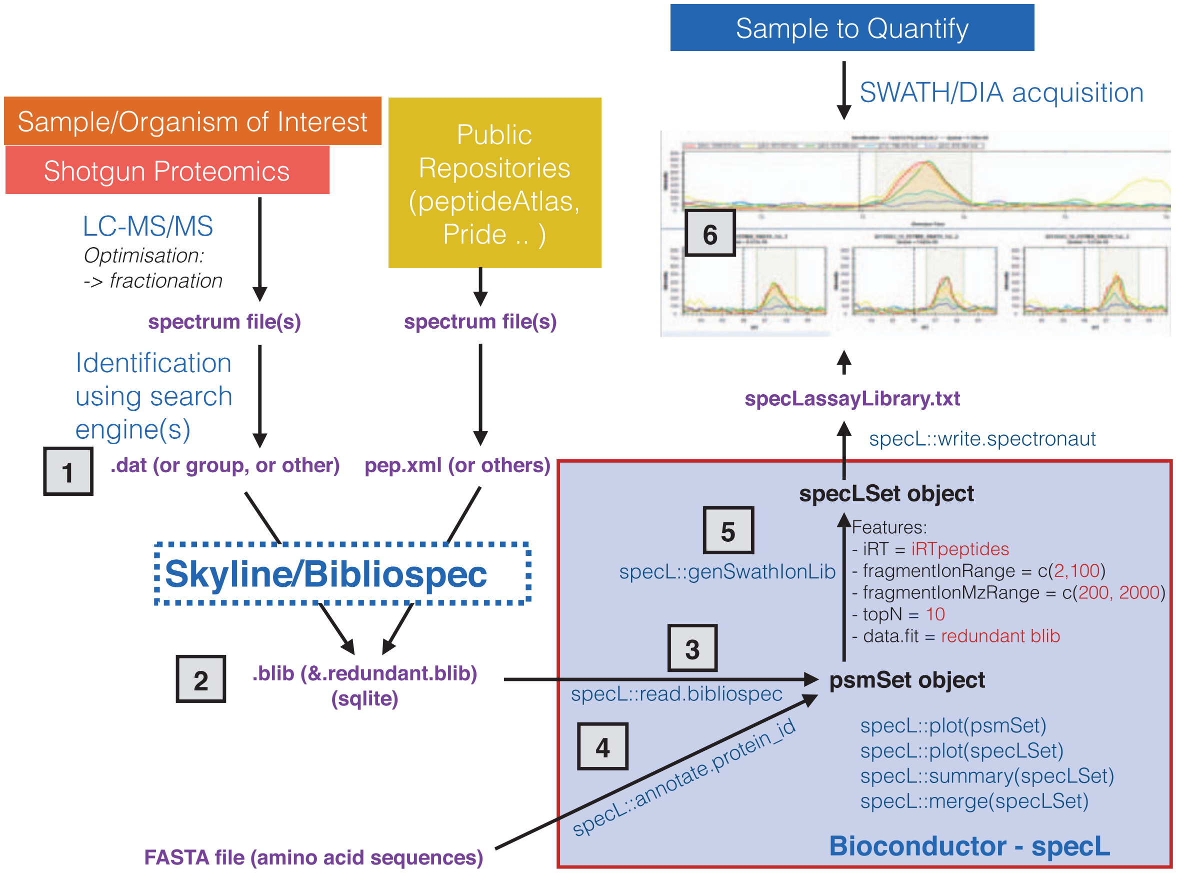

Fig. 1. A workflow overview for targeted data extraction using specL to generate an assay library is shown. In step 1, a peptide identification result is generated by a standard shotgun proteomics experiment or obtained from public repositories. Step 2 consists of building a spectral library with (MacLean *et al.*, 2010, Skyline). In Step 3, the read.bibliospec function reads the sqlite file from Skyline and generates a psmSet object. To re-annotate protein identification specL provides a function annotate.protein_id (Step 4) to associate protein identification along with the identified peptides. In Step 5, the genSwathIonLib function generates the specL object with a given set of parameters to build the customized spectral library. Step 6 is the export function to export the spectral library in the specific format which can be used in an external software to do the targeted data extraction, e.g. Spectronaut from BiognoSYS

representation of all precursors (m/z) plotted against their retention time. If the function is applied to a specL object four additional plots show the iRT normalization (with respective iRT peptides), a histogram of the independent retention times for all precursors as well as an *in silico* rt-fragment ion map and the distributions of fragment ion types (with respective charges) in the specL object. This should allow the identification of potential problems with the assay library or applied parameters.

4 Conclusion and outlook

This application note introduces the Bioconductor package specL, which reads the commonly used bibliospec spectral library format (blib from Skyline) and can process and visualize peptide spectrum matches thereof. specL is flexible enough to handle almost any search engine for which Skyline can build a blib file. Ultimately, it exports a text file-based assay panel, which is essential for targeted data extraction. To date, this step is either done using some in-house scripts which are rather un-structured and therefore cannot be easily used in other research groups or by limiting oneself to less flexible proprietary solutions provided by mass spectrometer vendors. The advantages of using specL are that it efficiently compiles all data in a consistent and reproducible way and therefore supports methods standardization. The specL package is flexible with respect to the peptide identification search engine results because it builds upon a Skyline file which supports most popular search engines. Also the assay panel can be customized (e.g. specifying the number of assays for each peptide or the type of fragment ions in the panel) and it features customizable retention time regression. In a future release of the specL package we intend to provide support for other export functions (e.g. TraML to support OpenSWATH or the PeakView format), also a merge function for specL sets and a consensus spectrum building function is foreseen.

Acknowledgements

The authors thank all colleagues of the Functional Genomics Center Zürich (FGCZ), and a special thank you goes to Simon Barkow-Oesterreicher, Peter Gehrig, Endre Laczko, Bertran Gerrits, Giancarlo Russo and Nathalie Selevsek for critical input and proof reading of the manuscript. We also thank the CRAN and Bioconductor team.

Funding

This work was supported by the EU 7th FP (Contract no. 262067—PRIME-XS and Contract no. 245143—TiMet).

Conflict of Interest: none declared.

References

Bernhardt,O.M. *et al.* (2012) Spectronaut: a fast and efficient algorithm for mrm-like processing of data independent acquisition (swath-ms) data. In:

Proceedings of the 60th American Society for Mass Spectrometry (ASMS) Conference on Mass Spectrometry, Vancouver, Canada, p. 68.

Editorial (2013) Method of the year 2012. *Nat. Methods,* **10**, 1.

Escher,C. *et al.* (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics,* **12**, 1111–1121.

Frewen, B. and MacCoss, M.J. (2007) Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinform.,* **13**, 13.7.

Gillet,L.C. *et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics,* **11**, O111.016717.

Lambert,J.P. *et al.* (2013) Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods,* **10**, 1239–1245.

MacLean,B. *et al.* (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* **26**, 966–968.

Morgan,M. *et al.* (2015) *BiocParallel: Bioconductor Facilities for Parallel Evaluation.* R package version 0.4.1.

Panse,C. *et al.* (2015) *protViz: Visualizing and Analyzing Mass Spectrometry Related Data in Proteomics.* R package version 0.2.11. http://cran.r-project.org/web/packages/protViz (5 March 2015, date last accessed).

Picotti,P. *et al.* (2013) A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature,* **494**, 266–270.

Rost,H.L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, **32**, 219–223.

Schubert,O.T. *et al.* (2013) The Mtb proteome library: a resource of assays to quantify the complete proteome of Mycobacterium tuberculosis. *Cell Host Microbe,* **13**, 602–612.

Schubert,O.T. *et al.* (2015) Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.,* **10**, 426–441.

Trachsel,C. *et al.* (2015) *specL: Prepare Peptide Spectrum Matches for Use in Targeted Proteomics.* R package version 1.1.13. http://www.bioconductor.org/packages/devel/bioc/html/specL.html (5 March 2015, date last accessed).