Using R and Bioconductor for proteomics data analysis

L. Gatto*,[1], L.M. Breckels[1], S. Gibb[2], A. Christoforou[1] and K. S. Lilley[1]

[1]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK
[2]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Germany
*lg390@cam.ac.uk – http://www.bio.cam.ac.uk/proteomics/

Introduction

R and Bioconductor have had a tremendous impact on the quality of genomics data analysis [6], demonstrating that extraction of relevant and biologically meaningful information from high-throughput data, required investing time and effort in the exploration and analysis of the data.

Even well-known and respected leaders in proteomics agree that it lies 10 years behind genomics. There are several valid reasons for this, including the chemical complexity of proteins, the technical complexity of the instrumentation (in particular mass-spectrometry - MS) and the vast possibilities in the study of proteins. An often overseen albeit essential component of this failure is arguably the software that is promoted inside the proteomics community. Computational proteomics researcher who value quality software, comprehensive data analysis and reproducible research ought to illustrate how more flexible and advanced tools can effectively be used and demonstrate their advantages. Here, we illustrate some examples of proteomics data analysis in R, in particular low level **raw MS data** manipulation, labelled and label-free **quantitation** and peptide **identification**, taken from the RforProteomics package [4].

Working with raw data

The proteomics community has developed a range of data standards and formats for MS data (the latest being mzML) to overcome the shortcomings or closed, binary vendor-specific formats. One of the main projects that implement parsers for the XML-based open formats is the C++ proteowizard project [2], which is interfaced by the mzR Bioconductor package using the Rcpp package.

```
library("mzR")
fname <- dir(system.file(package = "MSnbase", dir = "extdata"),
    full.name = TRUE, pattern = "mzXML$")
ms <- openMSfile(fname)
```

The resulting ms object is a file handle that allows fast direct and random access to the individual spectra. mzR is used by a variety of other packages like xcms, MSnbase, RMassBank and TargetSearch.

Challenges Further improve support of raw MS data and develop the range of supported formats, in particular identification (mzIdentML) and quantitation (mzQuantML) formats.

Labelled quantitation

The same raw data file can be imported in a convenient higher level container and directly processed, plotted, quantified and normalised with the MSnbase [5] software.

```
exp <- readMSData(fname, verbose = FALSE)
plot(exp[["X3.1"]], full = TRUE, reporters = iTRAQ4)
set <- quantify(exp, method = "trap", reporters = iTRAQ4,
    verbose = FALSE, parallel = TRUE)
head(exprs(set), n = 3)
```

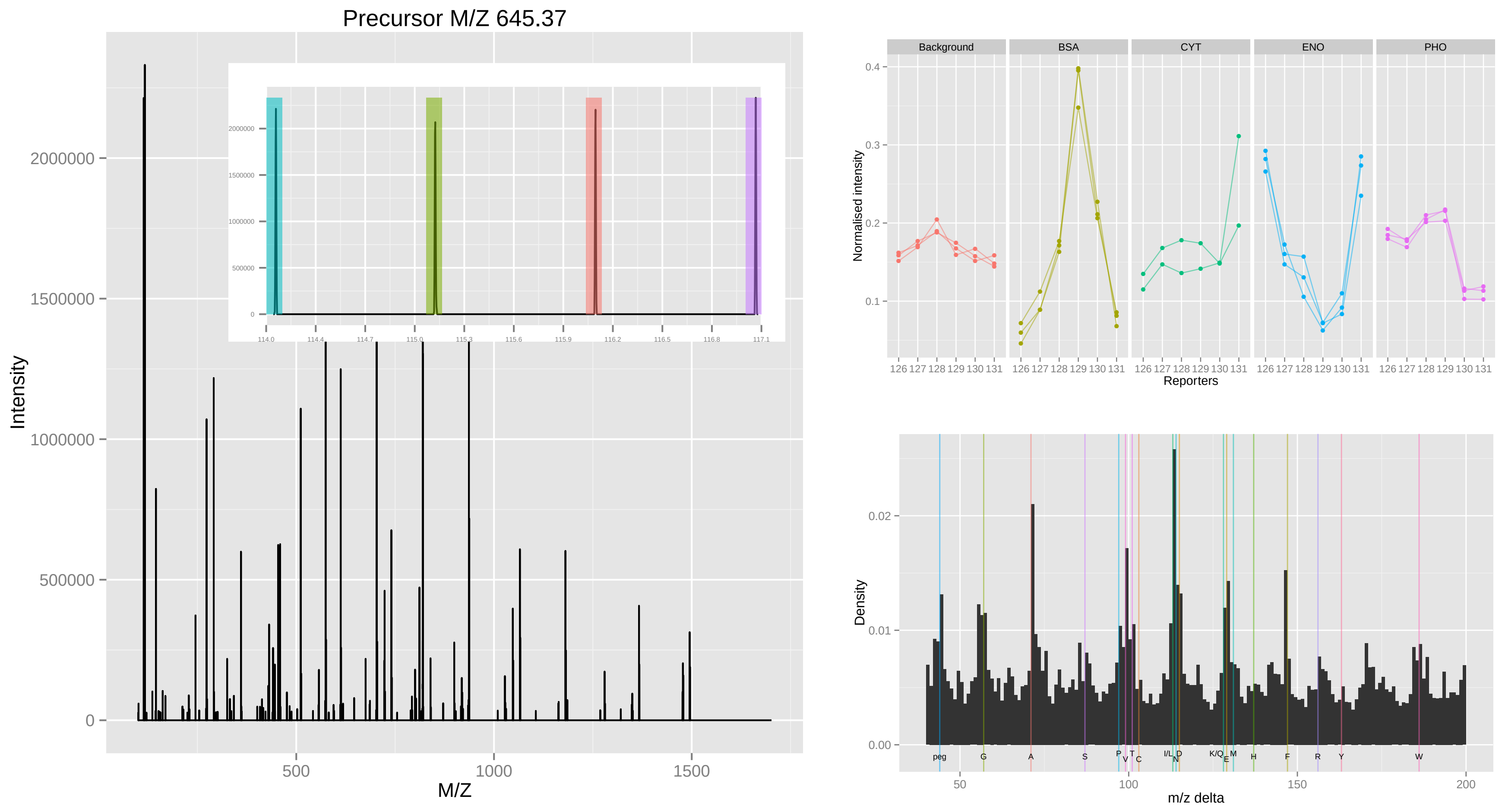

Figure: MS^2 spectrum of an iTRAQ 4-plex experiment showing the 4 isobaric reporter ions, as produced by plot above (left). Peptides of interesting from a spiked-in experiment (top right) and distribution of the m/z differences of all MS^2 spectra from the same experiment, use as a peptide-spectrum matching quality assessment (bottom right).

Challenges Although labelled MS^2 quantitation is well supported with MSnbase and isobar, metabolic labelling techniques like ^{15}N or SILAC still need to be developed.

Label-free quantitation

Label-free quantitation is available in the xcms [8] and MALDIquant [7] packages. The latter provides a complete pipeline, including baseline subtraction, smoothing, peak detection and alignment using warping functions, handling of replicated measurements as well as allowing spectra with different resolutions.

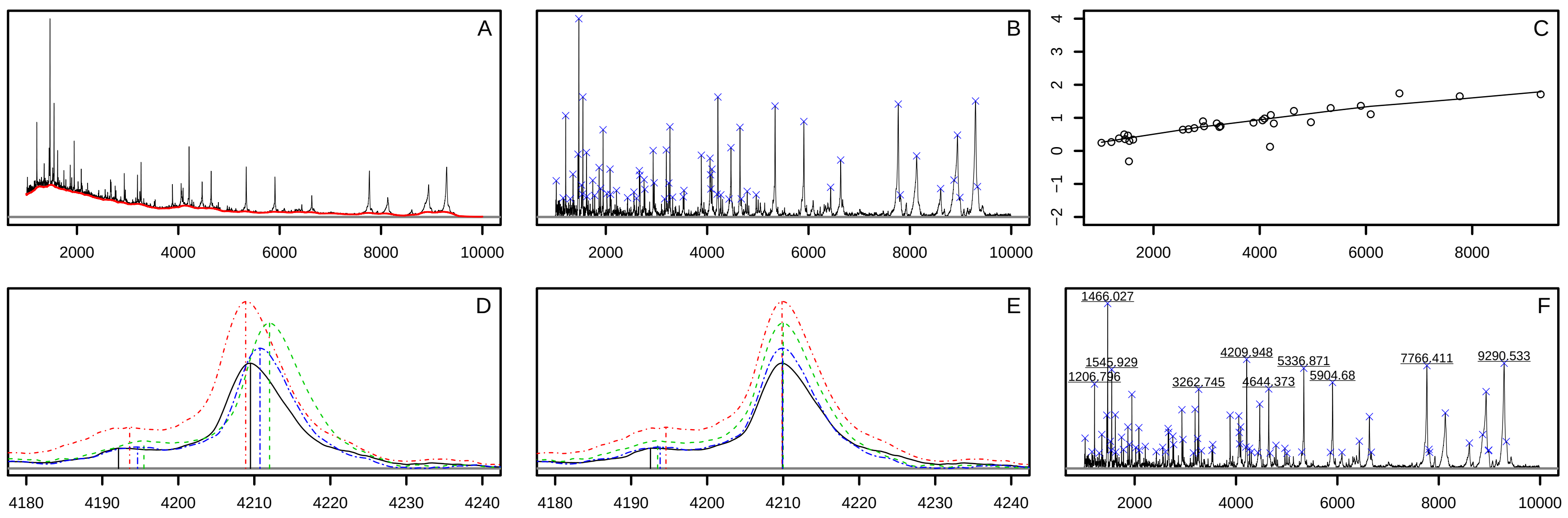

Figure: Illustration of the MALDIquant pipeline: raw spectrum with estimate baseline (A); variance-stabilized, smoothed, baseline-corrected spectrum with detected peaks (B); fitted warping function for peak alignment (C); four unaligned peaks (D); four aligned peaks (E); merged spectrum with discovered and labeled peaks (F).

A complete pipeline for MS^e data independent acquisition, including support for ion mobility separation is available in the synapter package [1] that, among other, transfers identification between acquisitions to substantially reduce missing values.

Challenges Application and benchmarking of label-free pipeline on popular Thermo Orbitrap instruments.

Peptide identification

The recently released rTANDEM package encapsulates the X!Tandem [3] search engine in R and uses the same XML-based parameter files as the native application. Result files can be directly parsed and mined in R .

```
xmlres <- rtandem(spectra.mgf, taxon = "yeast",
                  taxonomy = "taxonomy.xml",
                  default.parameters = "default-params.xml")
res <- GetResultsFromXML(xmlres)
proteins <- GetProteins(res)
peptides <- GetPeptides(res)
```

A complete pipeline with support for identification is welcome at the cost of additional development and maintenance time for developers. With support for mzIdentML files, it will become possible to import identification data from most search engines, this facilitating the integration of R based pipelines with existing tools.

Challenges Better integration of identification and raw/quantitation data infrastructure.

Conclusions and perspectives

The flexibility of the R environment and the breath of available packages is sometimes daunting for newcomers and introductory points of entry are welcome. The RforProteomics package [https://github.com/lgatto/RforProteomics] ought to assume such a role. For this, RforProteomics should be a collaborative project and contributions through the github repository are encouraged.

Despite well known advantages in terms of statistical analyses of data and some unique software for proteomics and mass-spectrometry data analysis, there remains a lot of efforts and work to be done for R/Bioconductor to become a complete framework for proteomics data processing. These efforts should be tackled by a group of developers. It is our hope that the RforProteomics will be a helpful targeted introduction to new users and motivate collaborative development of package developers.

[1] Bond NJ *et al. Improving Qualitative and Quantitative Performance for MS(E)-based Label-free Proteomics*. J Proteome Res. 2013 PMID: 23510225.

[2] Chambers M. *et al. A cross-platform toolkit for mass spectrometry and proteomics*. Nat Biotechnol. 2012 PMID: 23051804.

[3] Craig R and Beavis RC. *TANDEM: matching proteins with tandem mass spectra*. Bioinformatics. 2004 PMID: 14976030.

[4] Gatto L and Christoforou A. *Using R and Bioconductor for proteomics data analysis*. Biochim Biophys Acta. 2013 PMID: 23692960.

[5] Gatto L and Lilley KS *MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation*. Bioinformatics. 2012 PMID: 22113085.

[6] Gentleman R. *et al. Bioconductor: open software development for computational biology and bioinformatics*. Genome Biol. 2004 PMID: 15461798.

[7] Gibb S and Strimmer K *MALDIquant: a versatile R package for the analysis of mass spectrometry data.* Bioinformatics. 2012 PMID: 22796955.

[8] Smith CA *et al. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification*. Analytical Chemistry 2006 PMID: 16448051.

This work has been supported by the PRIME-XS project, grant agreement number 262067, funded by the European Union 7th Framework Program.