Controlling for confounding variables in MS-omics protocol: why modularity matters

Rob Smith, Dan Ventura and John T. Prince

Submitted: 24th May 2013; Received (in revised form): 17th June 2013

Abstract

As the field of bioinformatics research continues to grow, more and more novel techniques are proposed to meet new challenges and improvements upon solutions to long-standing problems. These include data processing techniques and wet lab protocol techniques. Although the literature is consistently thorough in experimental detail and variable-controlling rigor for wet lab protocol techniques, bioinformatics techniques tend to be less described and less controlled. As the validation or rejection of hypotheses rests on the experiment’s ability to isolate and measure a variable of interest, we urge the importance of reducing confounding variables in bioinformatics techniques during mass spectrometry experimentation.

Keywords: alignment; classification; peak picking; noise reduction; parameter selection

In science, we generate questions and design experiments to test possible answers to those questions. The ideal experiment involves a carefully controlled system where the impact of changes to a single variable may be measured. In practice, achieving full control over a system is difficult because the systems of most interest tend to be immensely complicated. Especially for complex systems, confounding variables—variables whose behavior can be spuriously attributed to the variable we are explicitly testing—may introduce hidden bias (referred to as omitted-variable bias), and therefore undermine an experiment. The degree to which omitted-variable bias is minimized is directly related to the accuracy of information, which may be gleaned from an experiment. The MS-omics (proteomics, lipidomics and metabolomics prosecuted via mass spectrometry) community has been extremely careful to control for confounding variables in sample preparation/processing and mass spectrometry analysis (hereafter called lab protocol). But somewhat surprisingly, this meticulousness has not extended as uniformly to data processing protocols in these same experiments. It seems obvious that data processing can and will influence experimental outcomes, just as changes to lab protocol do, and this influence should be expected to grow as the complexity of algorithms and data sets increases.

Consider two experiments carried out on the same mass spectrometer output files in search of drug biomarkers. In [1] and [2], the same group conducted two analyses of the same data set to predict cancer biomarkers. Despite the fact that the experimental variation was limited to the choice of post-processing bioinformatics tools (in this case, classification algorithms), the experiments yielded only two mutual m/z features (Figure 1). The diagnostic biomarkers selected as well as the sensitivity and specificity of the diagnostic changed solely due to the data processing protocol details (Table 1). Data processing protocol can and does influence experimental outcomes.

In data processing protocol, just as in lab protocol, limiting confounding variables boils down to limiting novel aspects under experiment. We suggest...
three guidelines to mitigate data-processing-related omitted-variable bias in MS-omics.

First, bioinformatics methods must be sufficiently described to permit replication, and parameters must be set to the established community standard independently demonstrated as effective in the literature. Although an explicit and careful protocol does not remove omitted-variable bias, it does make the potential confounding variables more obvious. It is hard to find an article that does not include meticulous lab protocol details: sample preparation and source, sample storage conditions, machine manufacturer and calibration settings and so forth. Unfortunately, when it comes to data processing, detailed descriptions are far too often replaced with descriptive snippets far too vague to reproduce the described protocol. In some articles, bioinformatics details are simply relegated to a flowchart box with a generic label like ‘data pre-processing’. No article that compressed all the details of the source, composition and preparation of a sample into a single flowchart box labeled ‘sample prep’ would ever pass peer review. Bioinformatics tools are unfortunately replete with free parameters that dramatically impact performance. If existing research suggests optimal parameter settings for a given situation, such settings should be used. If not, a reasonable search of the parameter space ought to be conducted and reported. A sub-optimal parameter setting can lead to lurking variable effects such as differential performance incorrectly attributed to the variable under experimentation. Although minimal reporting requirements suggested by HUPO-PSI (MIAPE, MIAMET, MIAME, etc.) are a step in the right direction, they do not require the reporting of all software parameters [3]. Consequently, an article can meet the HUPO-PSI minimum reporting standards and still be completely unreproducible.

Second, new bioinformatics tools or unproven parameter settings ought to be presented and evaluated independently from studies designed for clinical outcomes. It is already accepted that a new lab method deserves its own article in which there is sufficient room to describe the method in reproducible detail as well as to ascertain its strengths and weaknesses in controlled experiments over a variety of data sets. The same standard ought to apply to bioinformatics methods. All too often, an article whose focus is answering a chemical, biological or clinical question is used as a vehicle to present a novel data processing method. Introducing a new variable to study another variable should at the least be somewhat disconcerting to any scientist. It is far more clear and appropriate to present novel methods in their own right.

Third, and most importantly, whenever possible, bioinformatics algorithms ought to be implemented following the single responsibility principle—each module should have only one responsibility. In other words, algorithms ought to do one thing and do it well. This is not only a good programming philosophy but also a good experimental protocol philosophy that is at the heart of the scientific method—isolate and measure the variable of interest. New data processing methods, when coded modularly, are plug-in compatible with existing pipeline modules. Plug-in compatibility allows for a quick and comprehensive evaluation of new methods to ascertain downstream effects in the MS-omics pipeline. This approach has been implemented in frameworks for MS analysis (see, e.g. mzMine 2 [4] and OpenMS [5]), yet new algorithms are consistently presented independently of these frameworks. Not only are these new contributions more difficult to

Table 1: Comparison of cancer diagnostic biomarkers selected from two studies that used the same data but different post-processing tools

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of peaks</th>
<th>Diagnostic sensitivity (%)</th>
<th>Diagnostic specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam et al. [1]</td>
<td>9</td>
<td>83</td>
<td>97</td>
</tr>
<tr>
<td>Qu et al. [2]</td>
<td>12</td>
<td>92</td>
<td>95</td>
</tr>
</tbody>
</table>

Figure 1: A comparison of m/z values of biomarkers to diagnose cancer taken from two articles that used the same data set, with the only variability between them being the choice of bioinformatics tools used to post-process the data.
use and evaluate because of their independent packa-
ging, their non-modular interfaces and secondary
functionality (visualizations, data import/export,
etc.) are usually second rate to the full modular
frameworks mentioned earlier. Packaging an align-
ment algorithm with yet another two-dimensional
liquid chromatography–mass spectrometer display
makes about as much sense as bundling a newly
invented pipette with a second-rate centrifuge.
Modularity not only facilitates the isolation of new
variables but also decreases the learning curve by
cutting down the need to install new software,
learn new interfaces, deal with new file types
and facilitate transfer to the existing workflow.
Additionally, because of the number of confounding
variables and added obfuscation, lack of modularity
decreases the ease and transparency of evaluation
against other methods, stifling innovation and com-

We suggest that practitioners treat confounding
variables in the MS-omics tool kits with as much
care as they do with confounding variables in the
mass spectrometer experimental protocol. A modular
approach to bioinformatics tool development will help
minimize omitted-variable bias, make bioinformatics
tools interchangeable parts in the data processing pipe-
line and facilitate extensive evaluation in controlled
conditions before use in clinical application.

### Key Points

- Choice of bioinformatics data processing algorithms and param-
  eters affect the outcome of MS-omics experiments.
- Mitigation of confounding variables is just as important for data
  processing portions of the experiment as they are for lab por-
  tions and should be treated with as much care in practice and
detail in publication. Each data processing variable necessary to
reproduce the results ought to be reported in the article or sup-
plementary information, including software choices and param-
eter settings.
- Novel algorithmic protocols ought to be introduced in their own
dedicated article complete with either open-source code or suf-
icient detail to reproduce the algorithm as well as sufficient
evaluation with existing approaches to establish performance
and detail shortcomings. It is not appropriate to introduce
novel data processing techniques as a part of an experiment’s
protocol.

### FUNDING

National Science Foundation Graduate Research
Fellowship [DGE-0750759] (to R.S.).

### References

1. Adam BL, Qu Y, Davis JW, et al. Serum protein finger-
printing coupled with a pattern-matching algorithm distin-
guishes prostate cancer from benign prostate hyperplasia and
2. Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree ana-
lysis of surface-enhanced laser desorption/ionization mass
spectral serum profiles discriminates prostate cancer from
3. Taylor CF, Paton NW, Lilley KS, et al. The minimum
information about a proteomics experiment (MIAPE). *Nat
modular framework for processing, visualizing, and analyz-
ing mass spectrometry-based molecular profile data. *BMC
Bioinformatics* 2010;11:3.
5. Sturm M, Bertsch A, Gropl C, et al. Openms - an open-
source software framework for mass spectrometry. *BMC
tral.com/1471–2105/9/163.
6. Smith R, Ventura D, Prince JT. Novel algorithms and the
benefits of comparative validation. *Bioinformatics* 2013;