Check for updates

SOFTWARE TOOL ARTICLE

Authoring Bioconductor workflows with BiocWorkflowTools [version 1; referees: 2 approved with reservations]

Mike L. Smith 🆔 , Andrzej K. Oleś 🆔 , Wolfgang Huber 🆔

Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

Abstract

The Bioconductor Gateway on the F1000Research platform is a channel for peer-reviewed and citable publication of end-to-end data analysis workflows rooted in the Bioconductor ecosystem. In addition to the largely static journal publication, it is hoped that authors will also deposit their workflows as executable documents on Bioconductor, where the benefits of regular code testing and easy updating can be realized. Ideally these two endpoints would be met from a single source document. However, so far this has not been easy, due to lack of a technical solution that meets both the requirements of the F1000Research article submission format and the executable documents on Bioconductor.

Submission to the platform requires a LaTeX file, which many authors traditionally have produced by writing an Rnw document for *Sweave* or *knitr*. On the other hand, to produce the HTML rendering of the document hosted by Bioconductor, the most straightforward starting point is the R Markdown format. Tools such as *pandoc* enable conversion between many formats, but typically a high degree of manual intervention used to be required to satisfactorily handle aspects such as floating figures, cross-references, literature references, and author affiliations. The BiocWorkflowTools package aims to solve this problem by enabling authors to work with R Markdown right up until the moment they wish to submit to the platform.

Keywords
Bioconductor, workflow, reproducible research, markdown, scientific publishing

This article is included in the Bioconductor gateway.

Open Peer Review

Referee Status: ？ ？

	Invited Referees	
	1	**2**
version 1 published 06 Apr 2018	？ report	？ report

1 **Justin Kitzes**, University of Pittsburgh, USA

2 **Laurent Gatto** 🆔 , University of Cambridge, UK

Discuss this article

Comments (0)

Corresponding author: Mike L. Smith (mike.smith@embl.de)

Author roles: Smith ML: Conceptualization, Software, Writing – Original Draft Preparation; **Oleś AK**: Software, Writing – Original Draft Preparation; **Huber W**: Supervision, Writing – Review & Editing

Introduction

Bioconductor workflow vignettes are educational resources that demonstrate how one might tackle a particular multi-step bioinformatic analysis, primarily (but not necessarily exclusively) using the software found in the Bioconductor project[1]. They expand on the vignettes found in individual software packages by focusing on how multiple tools can be combined to conduct an analysis from beginning to end, rather than highlighting the features of a single resource. However they do share many similarities, in particular the desire to write such workflows in a literate programming style, with explanatory text surrounding executable code. This provides benefit to the reader, who can see each step of a workflow in context, and to the author, who can periodically check that the code is still valid and make changes to reflect either updates to the software they rely on, or improvements in methodology. These documents are then hosted on the [Bioconductor website] (www.bioconductor.org), which provides a centralized location for readers to find the articles and to download the software packages detailed within them. Workflow authors are encouraged to also submit their work as an article to F1000Research's Bioconductor Gateway, which provides the benefits (both to authors and readers) of increased visibility, peer-review and a citable reference. The intention is that (essentially) identical content will be present in both locations.

However, the requirements of the two publishing platforms are distinct. In order to regularly check code functionality and provide a workflow that is straight-forward to download and run by users, Bioconductor needs to be provided with documents written in R Markdown[2] or Sweave[3], which are compatible with the standard literate programming engines available for R. On the other hand, F1000Research request submissions in LaTeX or Microsoft Word format, where the code cannot be run directly. Both parties also apply their own style and branding to the final documents to present a coherent portfolio to end-users.

Given these distinct requirements, it has been somewhat difficult for an author to maintain a single document for submission to both platforms. This commonly results in prioritization of one over the other, followed by a non-trivial effort to convert to the other. Alternatively the author faces the challenge of writing two documents at the same time, trying to keep the information content synchronized, whilst dealing with two rather different syntaxes for document layout and formatting.

Here we present a strategy and accompanying tools to help authors develop and maintain a single document that can easily be transformed into the required format for submission to either platform.

Methods

Implementation

Given the intention for workflow documents to be full of executable examples that can be regularly checked and updated as necessary, it seems natural to recommend working with one of the literate programming formats available in R, rather than using a static typesetting tool. As previously mentioned, there are two formats commonly used here: Sweave and R Markdown. This immediately presents an author with a choice, even before a single word has been written, and there are reasonable arguments for electing to choose either; R Markdown has a simpler syntax and can be easily transformed into HTML for display on a website, while Sweave offers more precise control over document formatting and can readily be converted into a LaTeX format suitable for journal submission.

In order to streamline this, we have chosen to support only R Markdown as an input format, since this can be directly submitted to Bioconductor, with the conversion into the HTML format displayed on the website handled on their side. This then leaves the challenge of converting R Markdown into a format suitable for journal submission. To tackle this we have developed **BiocWorkflowTools**, an R package that provides article templates, conversion tools and the ability to upload documents to Overleaf.com (F1000Research's preferred LaTeX submission system).

Operation

In order to use **BiocWorkflowTools** the user must already have **R** version 3.4.0 or newer installed on their system. We also recommend working in the RStudio environment, however this is optional and all operations can be carried out at the command line with instructions for both approaches provided below.

Installation

BiocWorkflowTools can be obtained from the Bioconductor package repository by running the following commands in your R session.

```r
source("http://www.bioconductor.org/biocLite.R")
biocLite("BiocWorkflowTools")
```

Creating a Bioconductor workflow package

Given **BiocWorkflowTools's** raison d'être is to ease the burden of meeting the distinct requirements of two publishing platforms in a hassle-free manner as possible, our recommended strategy assumes that most authors begin a project with the intention of submitting the final outcome to both Bioconductor and F1000Research.

For F1000Research, the list of material required is straight-forward and familiar: the article itself, a list of references, figures, and supplementary materials. These can then be sent as a collection of files. When it comes to Bioconductor all the same materials are required; however their computing infrastructure, which enables the regular document checking and easy distribution, also requires that the submission is made in the form of an R package. There are numerous resources discussing how to create an R package[4] (and we would highly recommend potential authors to read these if they are not familiar with writing packages), but to streamline this process we provide the function `createBiocWorkflow`, which will create the minimum folder structure needed for submission to Bioconductor.

```
BiocWorkflowTools::createBiocWorkflow("MyWorkflow",
                                      quiet = TRUE,
                                      open = FALSE)
```

Running the example above will create a workflow package called *MyWorkflow* with the subdirectory *vignettes* containing an article template named *MyWorkflow.Rmd*. It is in this file that one should start developing their workflow document. In its initial state the template provides an exemplary skeleton of a typical workflow article, along with examples of how to include specific document features such as figures, tables, formulae and code blocks, in much the same way as the more traditional L^AT_EX and Microsoft Word templates available from F1000Research's website. The template also includes an example of the required document header, where article metadata including the title, author names, their affiliations and the abstract are specified.

In the example above, changing the argument `open = TRUE` will open a new RStudio project rooted in the newly created *MyWorkflow* folder.

Writing only a workflow document

If you do not wish to make a complete package, and instead would simply rather use R Markdown to author an F1000Research article, the recommended platform for most authors is still to work in RStudio. Rather than creating a new package as before, the user can opt to create a new R Markdown document from the file menu and, assuming the **BiocWorkflowTools** package has been installed, will be presented with the option to use the F1000Research Article template (Figure 1).

This will automatically open a new document based the F1000Research template described previously.

Working outside RStudio Even if you choose not to work in the RStudio environment, you can still use the included template to create a new file. We recommend using the template as a starting point to facilitate adherence to the required article structure. The command below will create a folder named *MyArticle* within the current working directory, and this in turn will contain the template **MyArticle.Rmd** which one can edit with the tool of choice.

```
rmd_file <- rmarkdown::draft("MyArticle.Rmd",
                             template = "f1000_article",
                             package = "BiocWorkflowTools",
                             edit = FALSE)
```

Conversion to L^AT_EX

If you are using RStudio to edit your workflow, the simplest way to create the L^AT_EX version of your document is to press the 'Knit' button above the document pane in your workspace. This will process the document and generate both the L^AT_EX version and a compiled PDF so you can see how the final print version will look. This process also carries out some necessary housekeeping, such as copying the required F1000Research L^AT_EX style file and the journal logo into the same location as your document, so you may notice some additional files appearing.

If one prefers to work in an editor other than RStudio, it is still possible to transform the R Markdown file into L^AT_EX (along with the aforementioned housekeeping) by using the function `render` from the **rmarkdown** package.
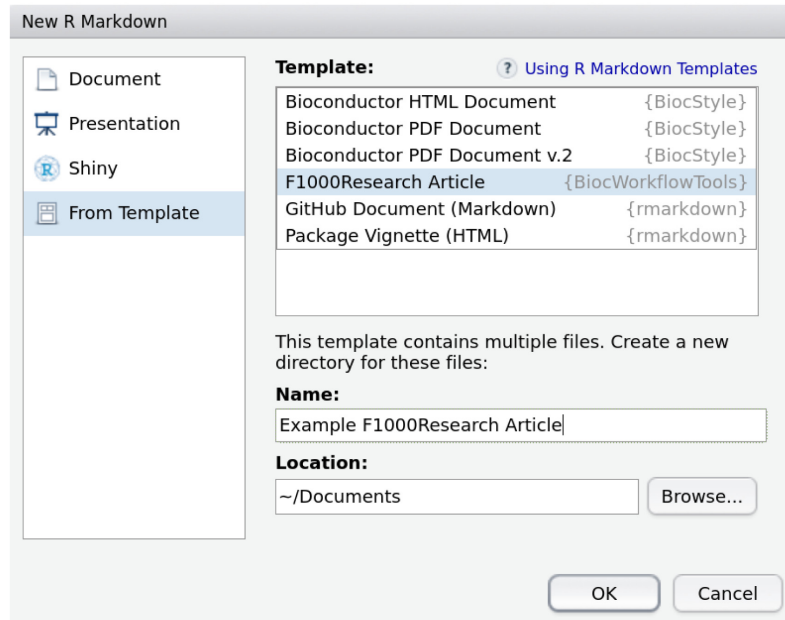
Figure 1. Creation of a new article is integrated into RStudio. The F1000Research template can be accessed via the 'New R Markdown' file menu dialog.

This will carry out much the same conversion process as using RStudio, executing the code chunks and producing the expected L^AT_EX and PDF output files.

```
rmarkdown::render(rmd_file)
```

Article submission

Ideally submission of a workflow is to both Bioconductor and F1000Research. Instructions for contributing to Bioconductor are available from https://www.bioconductor.org/developers/how-to/workflows.

Submission of L^AT_EX articles to F1000Research is currently performed using Overleaf.com, an online collaborative writing and publishing tool. Complete details are available online, but assuming one has already written an R Markdown workflow and generated a L^AT_EX source file, they can choose to upload the file directly to Overleaf using their web browser.

BiocWorkflowTools provides the function `uploadToOverleaf` as an alternative option for getting the article into the Overleaf system. This function takes the document directory and sends this to Overleaf, creating a new project for you automatically. The function will directly open the new project in a web browser.

```
BiocWorkflowTools::uploadToOverleaf("MyArticle")
```

BiocWorkflowTools::uploadToOverleaf("MyArticle")

At this point it is important to point out that both the L^AT_EX and R Markdown versions of the article are present in the Overleaf project, with the first of these being rendered into the document preview one sees on the site. In the Overleaf environment, only changes to the L^AT_EX version will be reflected in the preview pane, rather than the R Markdown the author has been working with until now. Thus it is easy for the two documents to become out of sync if edits are made using the browser interface. For this reason we recommend only working with Overleaf for the final submission process, and eschewing its document editing features at this point. Overleaf offers many attractive features such as collaborative editing between authors (particularly those who may not be familiar with the Rstudio environment) and live rendering of changes. However it should be emphasised again that it is natural to want to edit the L^AT_EX document in the Overleaf environment, which undermines the primary motivation behind **BiocWorkflowTools**, and as such our recommendation is to work exclusively on the Rmarkdown version of the article

(using whichever version control system and collaborative tools the authors prefer) and then use Overleaf purely as a submission tool.

Article revisions

Assuming the article is provisionally accepted, the journal editors will create a second, private, Overleaf repository for minor copy editing. Editorial comments and instructions will be included in the R Markdown document, enclosed in comment tags e.g. `<!-- Editorial comment -->`.

At this stage one can make changes to the R Markdown document in the web browser interface, however there is currently no way to regenerate the LaTeX containg the changes from here. Instead, to make additional changes after an Overleaf project has been created, we recommend utilising the fact that all Overleaf projects can be interfaced using git. Instructions for initialising this on your local machine provided by Overleaf[5]. Since this is a private project the author will need to supply their Overleaf username and password to use git. If the manuscript is already under git version control, the Overleaf server can be added as an additional remote repository. Authors can then work on the manuscript offline, and use the regular git commands to keep the local copy in sync, such as `git pull` to get the latest version from Overleaf. Before committing any edits made to the source Rmd file remember to also regenerate the LaTeX output file. Once they are ready to push the changes back to Overleaf, they will instantly appear online.

```
cd MyArticle
git pull
# make some changes and regenerate the Latex file
git commit -a -m "My edits"
git push
```

Once the author is happy they have addressed the editor's comments resubmission can be performed via Overleaf.

Summary

BiocWorkflowTools provides a straightforward set of helper functions and a document template discoverable in RStudio intended to simplify the process of authoring a source document that can both be run as an executable document, and also submitted to a journal for publication. We have focused on the collaboration between Bioconductor and F1000Research, and the authoring of R Markdown workflows. By supporting R Markdown and integrating the document template with RStudio, we hope we have made the path to meeting the requirements of both publication formats simpler for authors. Currently two distinct documents are still required for submission (R Markdown file, and the LaTeX file derived from this), but the potential exists to move the use of **BiocWorkflowTools** from the author to the journal, thus removing this final hurdle to the ambition of submitting a single source to both locations.

Although **BiocWorkflowTools** currently only supports the F1000Research article format, support for other scientific journals can be broadened in the future with the addition of more templates to the package, without requiring changes to the underlying workflow. The authors welcome approaches from publications that are interested in supporting the submission of R Markdown documents.

To demonstrate and test the tool's utility, this article has been written entirely in R Markdown.

Software availability

Software available from: http://bioconductor.org/packages/BiocWorkflowTools

Source code available from: https://github.com/grimbough/BiocWorkflowTools

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.1208607[6]

Software license: MIT

Competing interests

No competing interests were disclosed.

References

1. Huber W, Carey VJ, Gentleman R, *et al.*: **rchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Allaire JJ, Xie Y, McPherson J, *et al.*: **rmarkdown: Dynamic Documents for R.** R package version 1.7. 2017.
    Reference Source

3. Leisch F: **Sweave. dynamic generation of statistical reports using literate data analysis.** 2002.
    Reference Source

4. Wickham H: **R packages: organize, test, document, and share your code.** "O'Reilly Media, Inc.", 2015.
    Reference Source

5. **Overleaf.com.** Date Accessed: 21-03-2018.
    Reference Source

6. Smith ML, Oleś AK: **grimbough/BiocWorkflowTools: F1000 Article Submission (Version F1000).** *Zenodo.* 2018.
    Data Source

Open Peer Review

Current Referee Status: ❓ ❓

❓ **Laurent Gatto** (iD)

Cambridge Centre for Proteomics, University of Cambridge, Cambridge, UK

The authors describe the BiocWorkflowTools Bioconductor package, a tool to write and maintain a single entry point that can be converted and submitted as a Bioconductor workflow or as a F1000Research Bioconductor gateway article.

My understanding is that the authors effectively want to address or facilitate three challenges with their packages:

1) The technical challenge that is allowing authors to use R markdown to generate outputs that will fit the requirements set by Bioconductor (for the submission of an R package containing a workflow vignette) and F1000Research (submission of a LaTeX document). This purely technical goal is elegantly achieved thanks to the BiocWorkflowTools package that leverages existing infrastructure and templates.

2) The authors also seem to favour submission of workflows to both Bioconductor and F1000Research, so that authors and readers will benefit from the advantages of both platforms, i.e. reproducibility from the former, and visibility and citable reference from the latter.

However, here I feel that the solution falls short because these two options rely on two pipelines in BiocWorkflowTools, namely *BiocWorkflowTools::createBiocWorkflow* for Bioconductor workflows and *rmarkdown::draft(template = "f1000_article")* for an F1000Research article.

The problem I see is that this will essentially define from the beginning what distribution platform future workflow authors will favour, instead of being able to write the workflow with BiocWorkflowTools and then submit to both locations.

Technically, the article corresponds to the vignette of the R package. The package also provides additional meta-data in the DESCRIPTION file. It would be easy for a user that has experience in package development to convert one in the other, but I am wondering if a single pipeline wouldn't be a better option for authors that don't have any experience in package development. Specifically, I am thinking of either converting the Rmd article to a package (by populating the DESCRIPTION file using meta-data in the article's header), or choosing the package as single pipeline and uploading the vignette directory using *BiocWorkflowTools::uploadToOverleaf*.

3) The last issue, arguably the most difficult one to address, is to bridge the gap between the fully reproducible scientific document and the manual publishing pipeline.

Ironically, the greatest breakthrough here seems to be that *"Editorial comments and instructions will be included in the R Markdown document"* using Overleaf as an intermediate *broker*. To avoid breaking the reproducible pipeline, authors are advised to use the Overleaf project (which is also a git repository) as a remote and pull changes into their local repository.

Establishing a pipeline that maintains reproducibility and traceability beyond submission is very important. While it isn't strictly part of the tool, it would be valuable to describe how to set this up in a bit more details, or provide links to useful resources.

Another dimension that could be documented is when workflow authors already work collaboratively on the article/package using a git repository via, for example, GitHub. They would have their local/remote repositories for the duration of the writing and, when ready to submit, upload to an Overleaf project (that itself is a remote git repository). Editorial comments are then returned in a new Overleaf project/github remote repository. Authors would then add one (or two) additional remote repositories pointing to the Overleaf project(s). Finally, it would be useful to also describe how a revision of the article would be updated and resubmitted via Overleaf.

My suggestions for this manuscript are two-fold:

1. Comment on the usefulness of having two independent workflow pipelines (one through a package for Bioconductor, and another one through an Rmd file for F1000Research) or a single one. I fear that having two will limit the submission to both locations.

2. Even though it isn't directly related to the BiocWorkflowTools package, it would be very useful for the authors to provide more details and/or links to relevant resources to further integrate the local workflow repository to the remote Overleaf project to maintain reproducibility beyond the first submission.

Minor comments:

- The first letter of 'Orchestrating' is missing in the title of Ref. 1.
- The link to Bioconductor in the first paragraph is mis-formatted.
- The code in the code chunk describing how to upload the article to Overleaf is repeated.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 26 April 2018

Justin Kitzes
University of Pittsburgh, Pittsburgh, PA, USA

This article describes a new package, BiocWorkflowTools, that is designed to allow authors to write a manuscript in a single R Markdown file and then easily submit the manuscript to both F1000Research and Bioconductor, two platforms with different submission requirements. The goal is to reduce the risk that versions of the manuscript published in both places will begin to inadvertently diverge. The package addresses an important and somewhat general problem in scientific publishing, which is how to keep multiple versions of a published article in sync through the revision and publication process.

The article is well-written and describes the use of BiocWorkflowTools clearly. I have several major and minor comments on the article. These mainly request clarification or expansion of the points made in the manuscript, and none should stand in the way of the article's acceptance.

Major comments

- As the article is brief, it does not provide instructions for the many associated steps that could be necessary to use the package, such as how to structure an R package, set up an Overleaf account, write in Markdown, write in LaTeX, and use git (both locally and remotely). It would seem reasonable, though, to assume that the target audience for the package will already have these skills.

- While I understand the problem that the package is attempting to solve, I might also question the wisdom of a publication strategy that requires this package at all. That is, I question whether it's a good idea, even with this package, to attempt to duplicate nearly identical content across two platforms. I can see the value added by having an article published on the F1000 platform, but is the value of being able to execute the embedded code blocks really worth the trouble of duplicating the entire article on Bioconductor?

  As an alternative, although it would take more effort, what about having Bioconductor scrape the HTML from F1000 (or finding a way to access the article text in a structured format) and extract code blocks when the manuscript occasionally needs to be executed? This would add burden on the Bioconductor side but would allow authors to only worry about a single public document.

- I'm curious as to why the package developers chose to go through LaTeX and Overleaf for F1000 submission rather than the Word route (e.g., by converting R Markdown to Word and then having authors submit the Word version to F1000). I would be slightly concerned that involving another third party (Overleaf) increases the potential for the package breaking in the future due to changes to the Overleaf API, Overleaf's default package list, the LaTeX packages themselves, F1000's relationship with Overleaf, etc. In comparison, the DOC/DOCX format is relatively stable. I was more enthusiastic about the Overleaf workflow until reading all of the caveats and warnings surrounding the appropriate approach to making revisions in Overleaf, and how that can cause the very problems the package is designed to prevent.

- Would it be possible to add a function to the package that would automatically check for synchronization between the F1000 and Bioconductor versions of an article (this would likely involve web scraping as mentioned above)? If the two versions did go out of sync, it will likely be very difficult to detect that divergence.

- Does the package correctly handle all of the issues described in the abstract in the conversion from Markdown to LaTeX, including "floating figures, cross-references, literature references, and author affiliations"?

- How does this package work under the hood? In particular, is pandoc a dependency and can this package be understood as a kind of pandoc wrapper that handles the special cases listed above?

Minor comments

- There is an incorrectly written Markdown link that is currently appearing as "[Bioconductor website] (www.bioconductor.org)" - I believe the leading "http://" will be needed for this to render properly as a link.

- Awkward wording - "Ideally submission of a workflow is to both Bioconductor and F1000Research" could be "Ideally a workflow is submitted to both…"

- There appears to be a duplicated line that appears both inside and outside of a code block: " *BiocWorkflowTools::uploadToOverleaf("MyArticle")*"

- Do I understand correctly that the F1000 editors will make comments in an R markdown file that appears on Overleaf, but then expect revisions to occur in LaTeX? This seems to be an unusual setup, just confirming that it's correct.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research