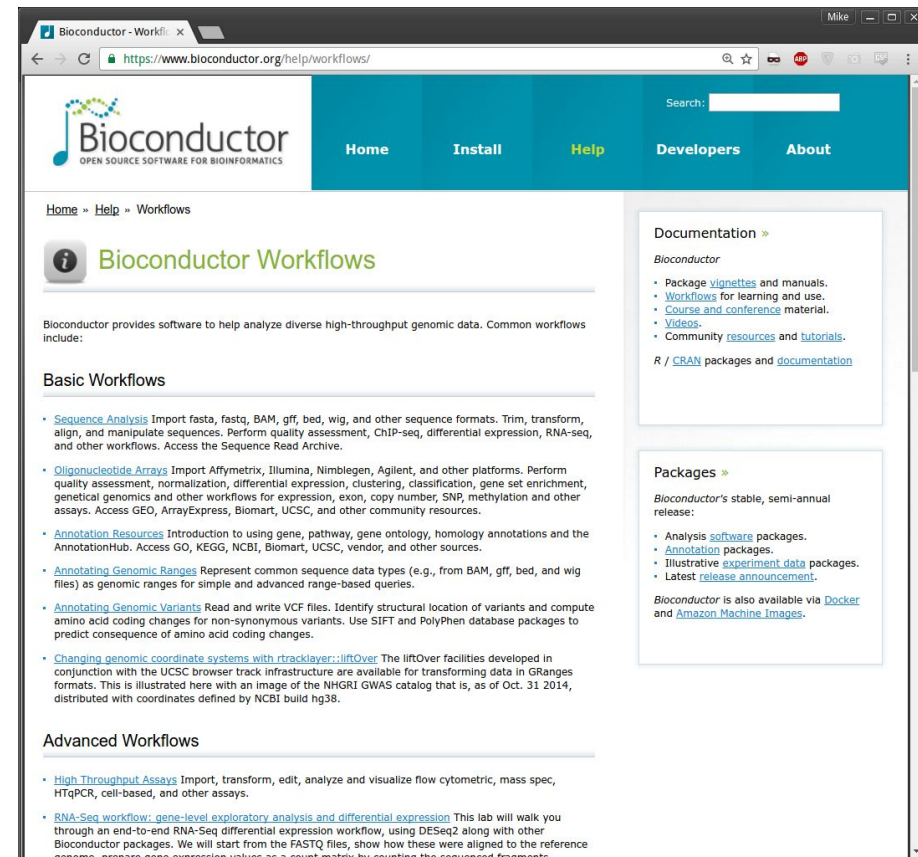


Publishing Live Analysis Workflow Documents with Bioconductor and F1000Research

Mike Smith, Andrzej Oleś, Wolfgang Huber

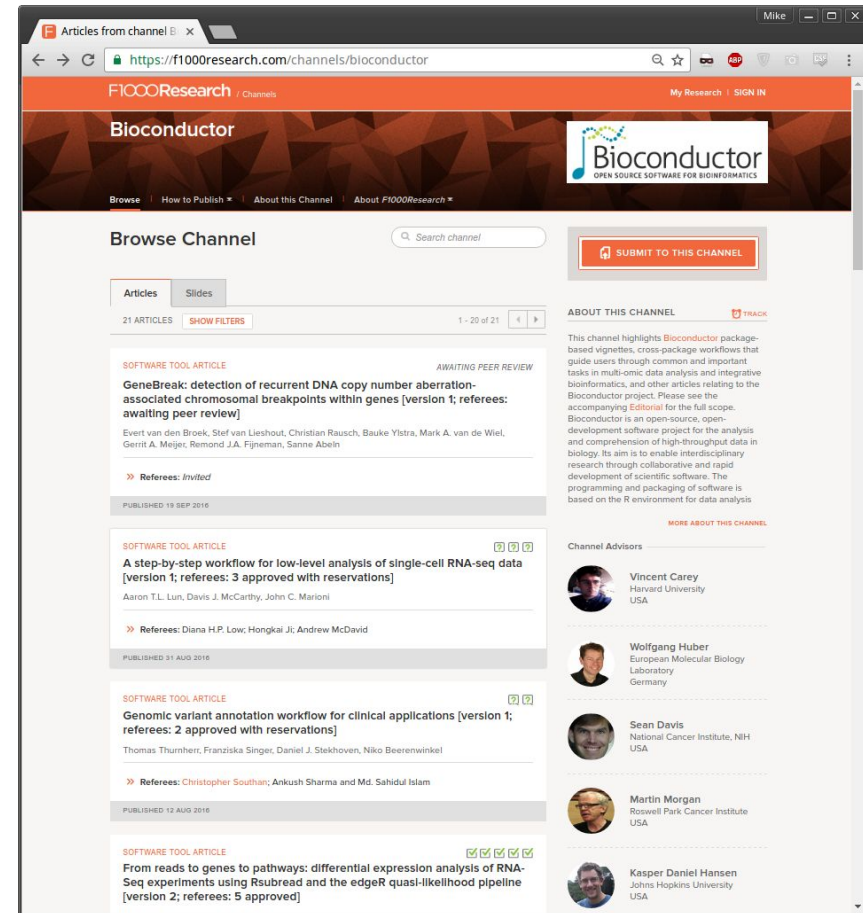
Bioconductor Workflows

- Educational documents detailing how to use multiple packages
- Regularly checked on Bioconductor servers
 - Ensures changes in component tools are identified quickly
 - Allows rapid updates by author
 - Provides users a platform to easily access the complete suite of tools



Bioconductor & F1000 Research

- Difficult to get credit
- F1000Research provides a citable, peer-reviewed publication platform
 - Currently 21 submitted articles
 - Combined they have > 37,000 views & > 9,000 downloads
- Intention is for the same document to be submitted to both platforms - updates are possible



Summer 2016

- Call for papers to coincide with BioC Conference
- 8 papers submitted to F1000, but not to BioC
- Authors prioritise the journal submission over BioC website
- Emailed authors and offered help getting BioC workflow implemented

‘Survey’ of authors’ methodology

- Four authors responded to our offer of assistance
- Variety of approaches to produce their workflows
 - Rmd, data in hardcoded local paths, not publicly available
 - Rmd (code only), text in LaTeX, data from GEO / institute website
 - Rmd, all code duplicated (EVAL once, ECHO once), data from TCGA
 - Sweave, data from GEO
- Perhaps some clearer guidelines would help?

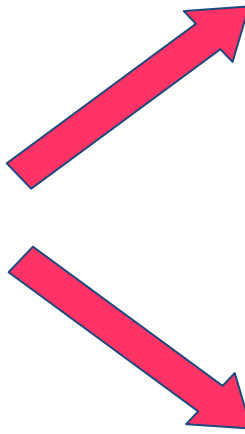
Refined author guidelines

- Start from Rmarkdown
 - HTML for Bioconductor
 - Reduce number of BioC build tasks
 - Simpler to write instructions for one format
- Create package for each workflow
 - Small data set easily included
 - Larger data can be a data package, or external (maybe not 'personal' website)
- Clearer examples of how to include/access example dataset

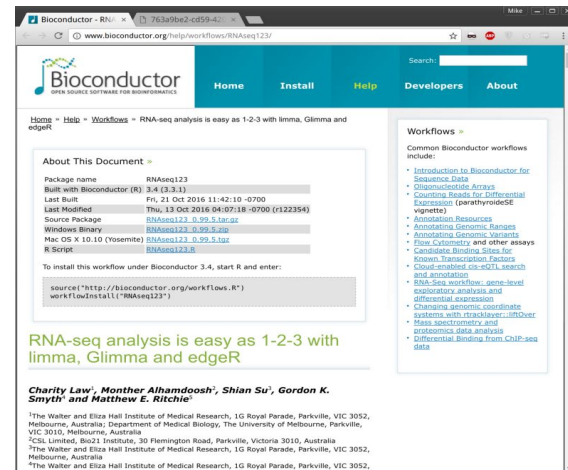
Idealised authoring process

Source R Markdown

```
1 ---
2 title: "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR"
3 author:
4   - name: Charity Law
5     affiliation: The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade,
6     Parkville, VIC 3052, Melbourne, Australia; Department of Medical Biology, The University of
7     Melbourne, Parkville, VIC 3010, Melbourne, Australia
8   - name: Monther Alhamdoosh
9     affiliation: CSL Limited, B1021 Institute, 30 Flemington Road, Parkville, Victoria 3010,
10    Australia
11 date: 18 September 2016
12 vignette: >
13   %\VignetteIndexEntry{RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR}
14   %\VignetteEngine{knitr::rmarkdown}
15   %\VignetteEncoding{UTF-8}
16 bibliography: workflow.bib
17 output: workflow.bib
18 blockstyle: latex_document
19 fig_caption: true
20 self_contained: no
21 ---
22 # Abstract
23 The ability to easily and efficiently analyse RNA-sequencing data is a key strength of the
24 Bioconductor project.
25 Starting with counts summarised at the gene-level, a typical analysis involves pre-processing,
26 exploratory data analysis, differential expression testing and pathway analysis with the results
27 obtained informing future experiments and validation studies.
28 In this workflow article, we analyse RNA-sequencing data from the mouse mammary gland,
29 demonstrating use of the popular rnaSeq package to import, organise, filter and normalise the
30 data, followed by the limma package with its voom method, linear modelling and empirical
31 Bayes moderation to assess differential expression and perform gene set testing. This pipeline is
32 further enhanced by the glimma package which enables interactive exploration of the results so
33 that the raw counts from an RNA-sequencing experiment into biological
34 insights using Bioconductor.
```



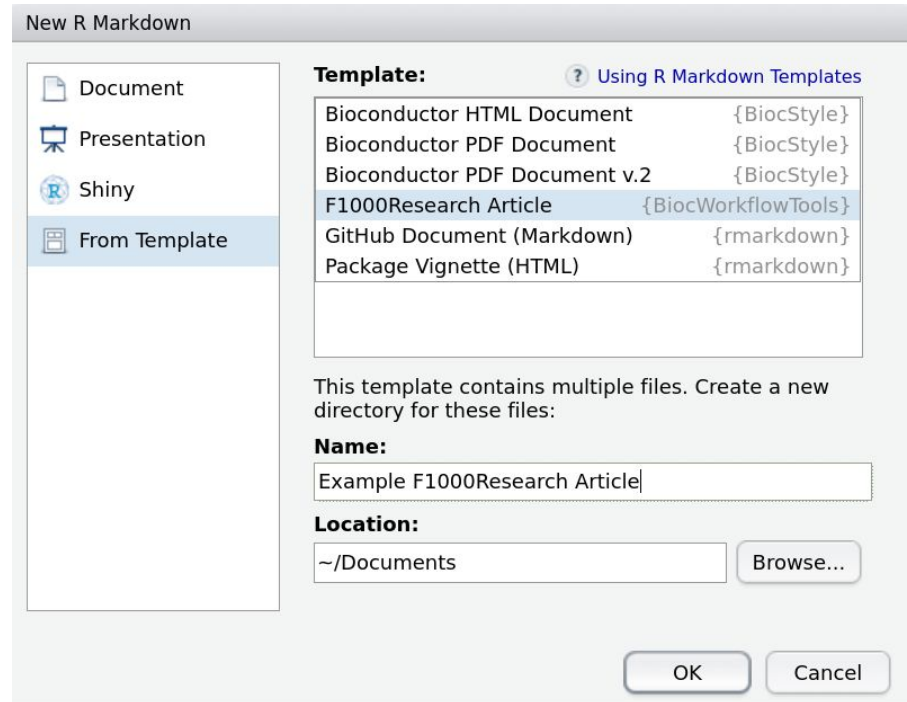
F1000Research PDF



Bioconductor HTML

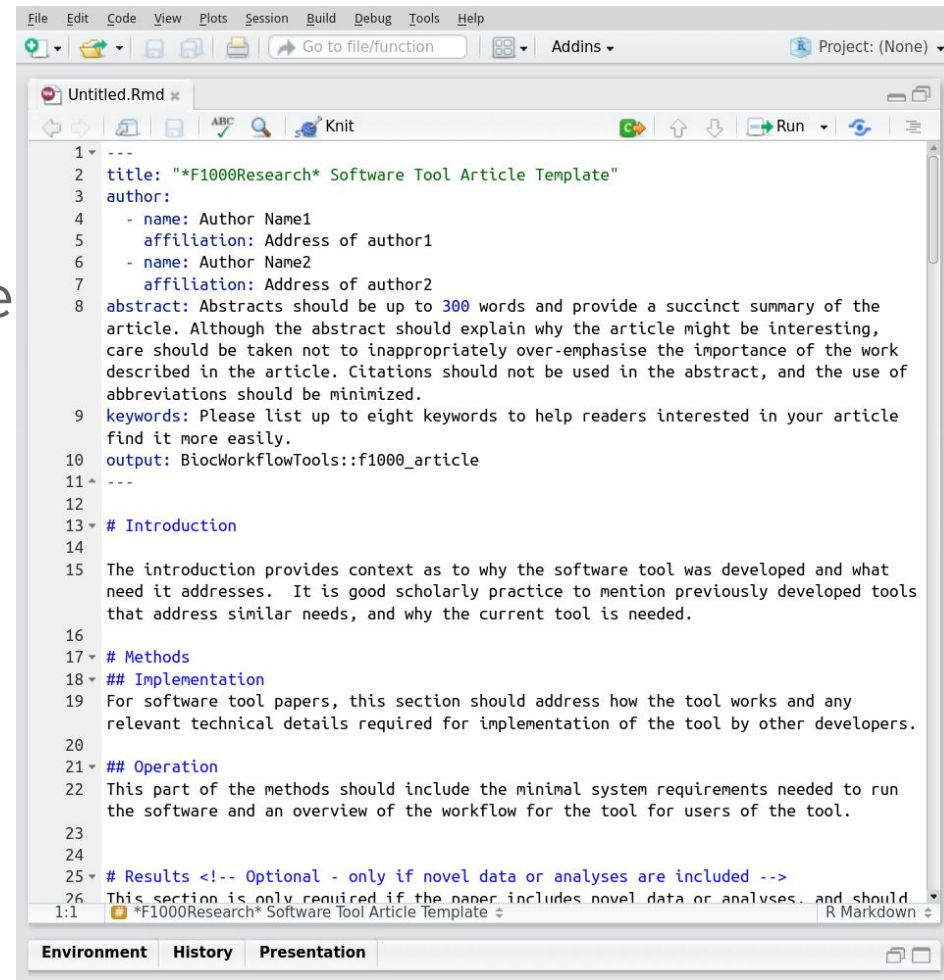
BiocWorkflowTools Package

- Started as a conversion function between RMarkdown and LaTeX
- Based on knitr and pandoc
- Now has RStudio integration to start a new article



Article Template

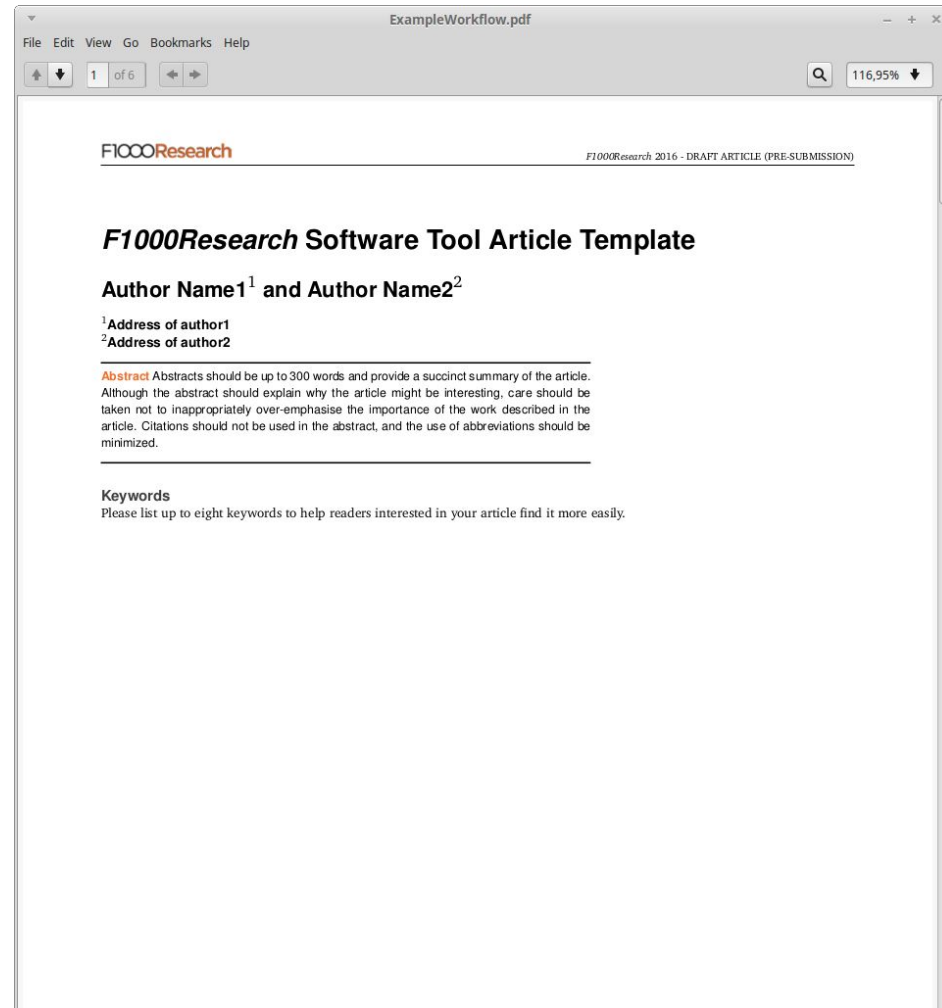
- Provides a template, based on F1000 LaTeX template
- Provides document structure and markdown examples for:
 - Code integration
 - Math & formulas
 - Tables
 - Figures, labelling & referencing
 - Citations



```
1 ---
2 title: "*F1000Research* Software Tool Article Template"
3 author:
4   - name: Author Name1
5     affiliation: Address of author1
6   - name: Author Name2
7     affiliation: Address of author2
8 abstract: Abstracts should be up to 300 words and provide a succinct summary of the
9 article. Although the abstract should explain why the article might be interesting,
10 care should be taken not to inappropriately over-emphasise the importance of the work
11 described in the article. Citations should not be used in the abstract, and the use of
12 abbreviations should be minimized.
13 keywords: Please list up to eight keywords to help readers interested in your article
14 find it more easily.
15 output: BiocWorkflowTools::f1000_article
16 ---
17 # Introduction
18 The introduction provides context as to why the software tool was developed and what
19 need it addresses. It is good scholarly practice to mention previously developed tools
20 that address similar needs, and why the current tool is needed.
21 # Methods
22 ## Implementation
23 For software tool papers, this section should address how the tool works and any
24 relevant technical details required for implementation of the tool by other developers.
25 ## Operation
26 This part of the methods should include the minimal system requirements needed to run
27 the software and an overview of the workflow for the tool for users of the tool.
28 # Results <!-- Optional - only if novel data or analyses are included -->
29 This section is only required if the paper includes novel data or analyses... and should
30 *F1000Research* Software Tool Article Template
```

Compiling an article

- Running `knit()` produces F1000 themed PDF
- Retains LaTeX source for submission
- Copies necessary images and styles from `BiocWorkflowTools` to working directory



Uploading to Overleaf

- F1000Research submission is done via Overleaf.com
- `uploadToOverleaf()`
 - zips a folder
 - Pushes to Overleaf and creates new project
 - (Optionally) opens browser at this location
- Can now submit to the journal
- Every Overleaf project is also a git repo, so further changes can be committed and pushed

Caveats

- Work with one document upto the point of submission - what happens if editors make changes?
- Hopefully code remains untouched
- Currently working on a tools to match text blocks between LaTeX and RMarkdown docs so changes can be integrated back into the source
- Definitely still a work in progress

Acknowledgements

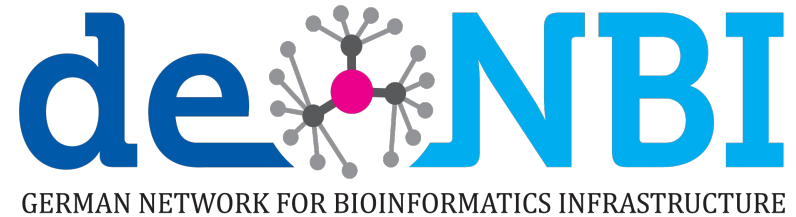
Andrzej Oleś

Wolfgang Huber

Martin Morgan

Mike Love

Thomas Ingraham



Generate F1000 LaTeX from Rmd

- `knit()` evaluates the code chunks
 - Have to force LaTeX output from Rmd input
 - Generated file is a 'half-way house' - Markdown with LaTeX code chunks
- `pandoc_convert()` produces LaTeX file
 - Pass template file with F1000 style parameters
 - Turn off word wrap!
 - F1000 allow many citation styles, we can provide them here