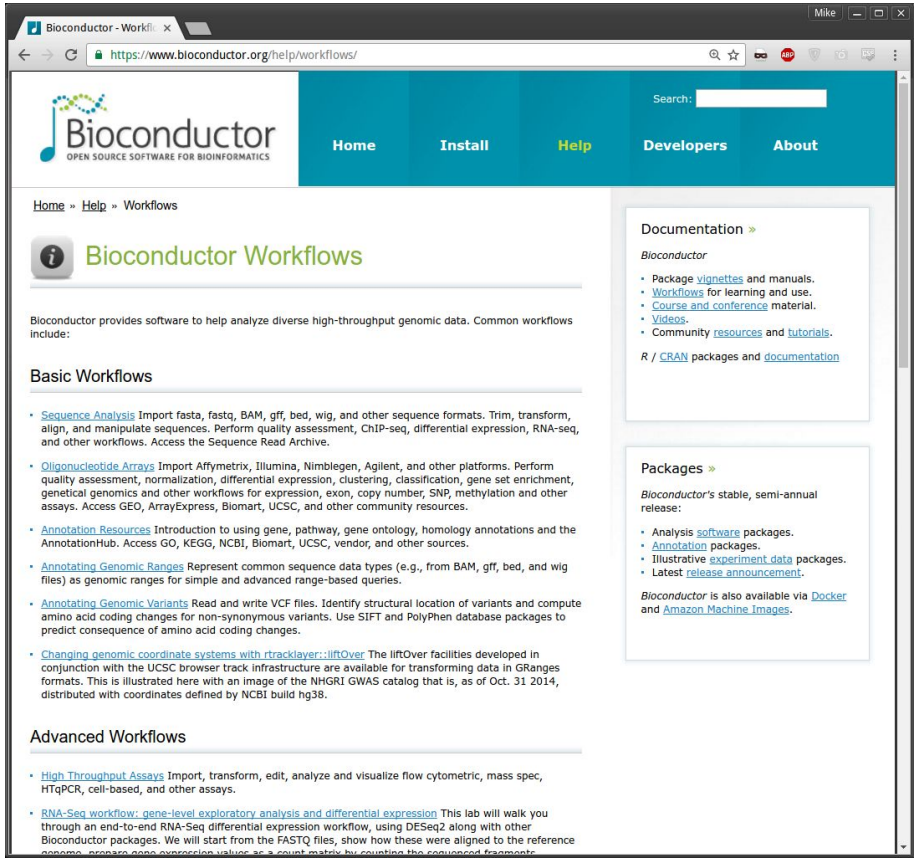
Publishing Live Analysis Workflow Documents with Bioconductor and F1000Research

Mike Smith, Andrzej Oleś, Wolfgang Huber

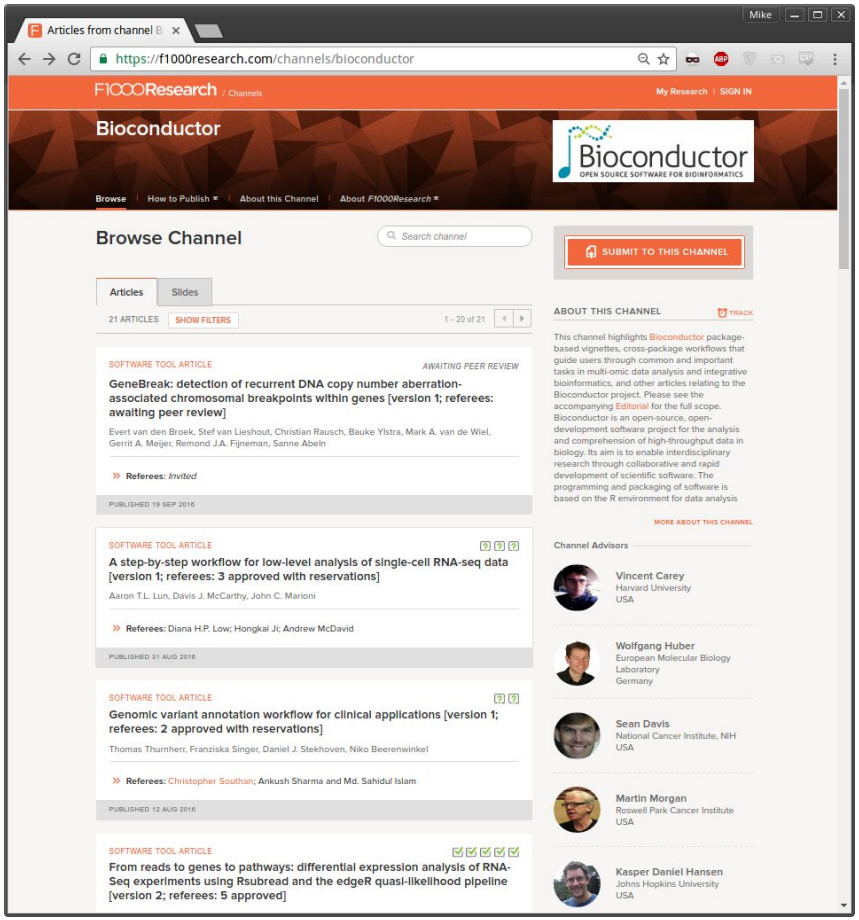
Bioconductor Workflows

- 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

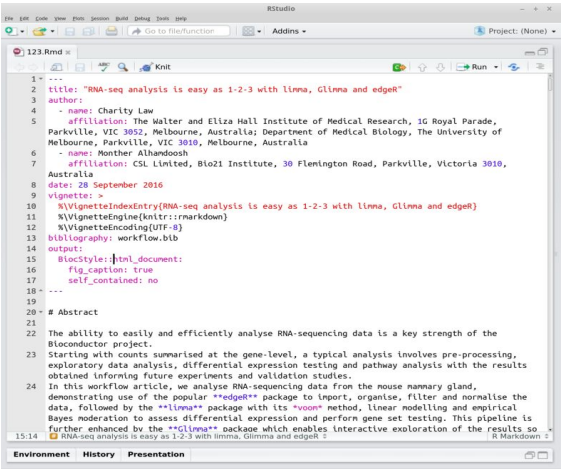
- Four authors responded to our offer of assistance
- Variety of approaches to produce their workflows
  - Rmd, data in hardcoded local paths, not publicly available
  - Rmd (code only), text in LaTeX, data from GEO / institute website
  - Rmd, all code duplicated (EVAL once, ECHO once), data from TCGA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

Refined author guidelines

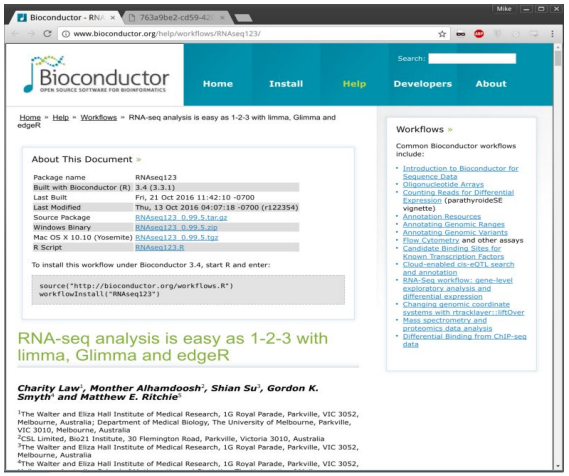
- Start from Rmarkdown
  - HTML for Bioconductor
  - Reduce number of BioC build tasks
  - Simpler to write instructions for one format
- Create package for each workflow
  - Small data set easily included
  - Larger data can be a data package, or external (maybe not ‘personal’ website)
- Clearer examples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

Idealised authoring process

Source
R Markdown



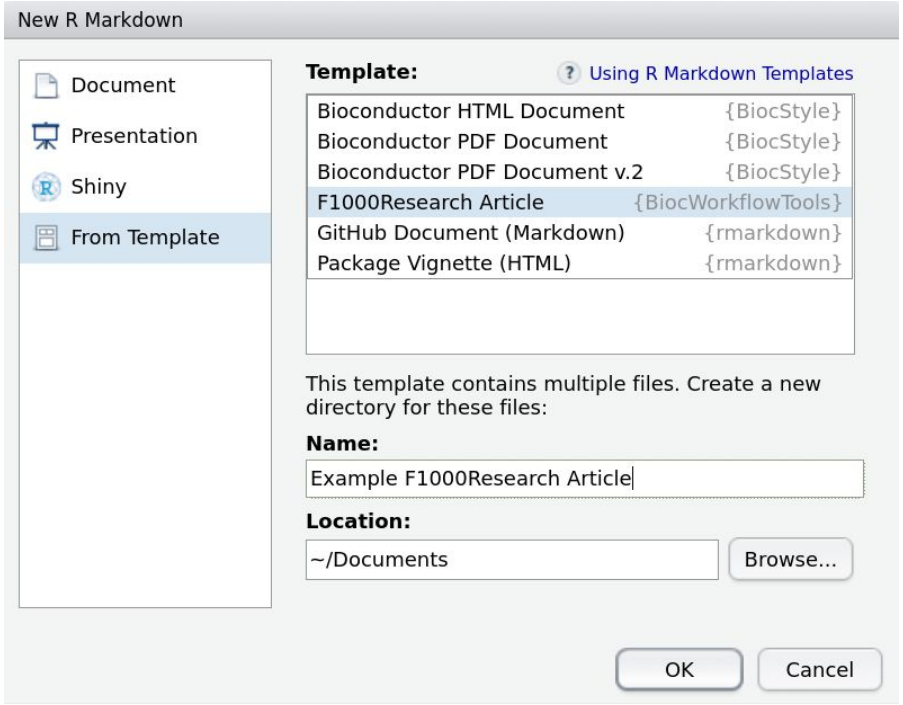
F1000Research
PDF



Bioconductor
HTML

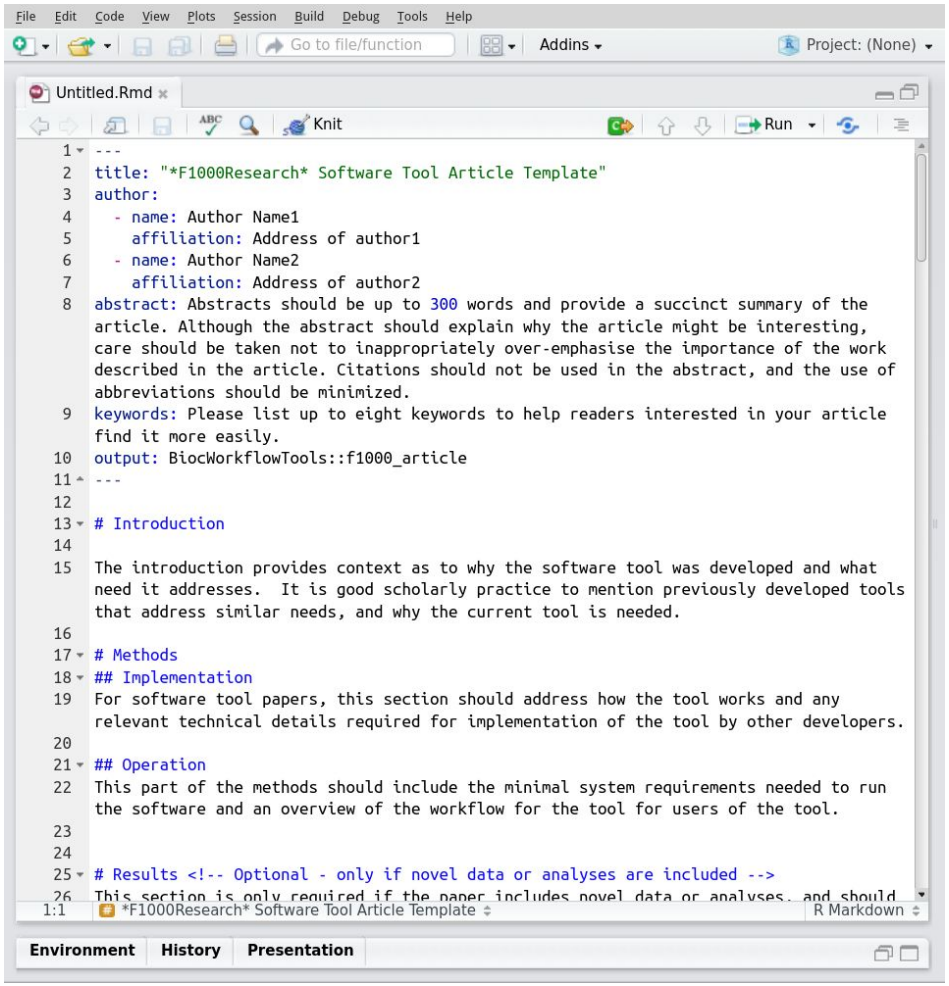
`BiocWorkflowTools` Package

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



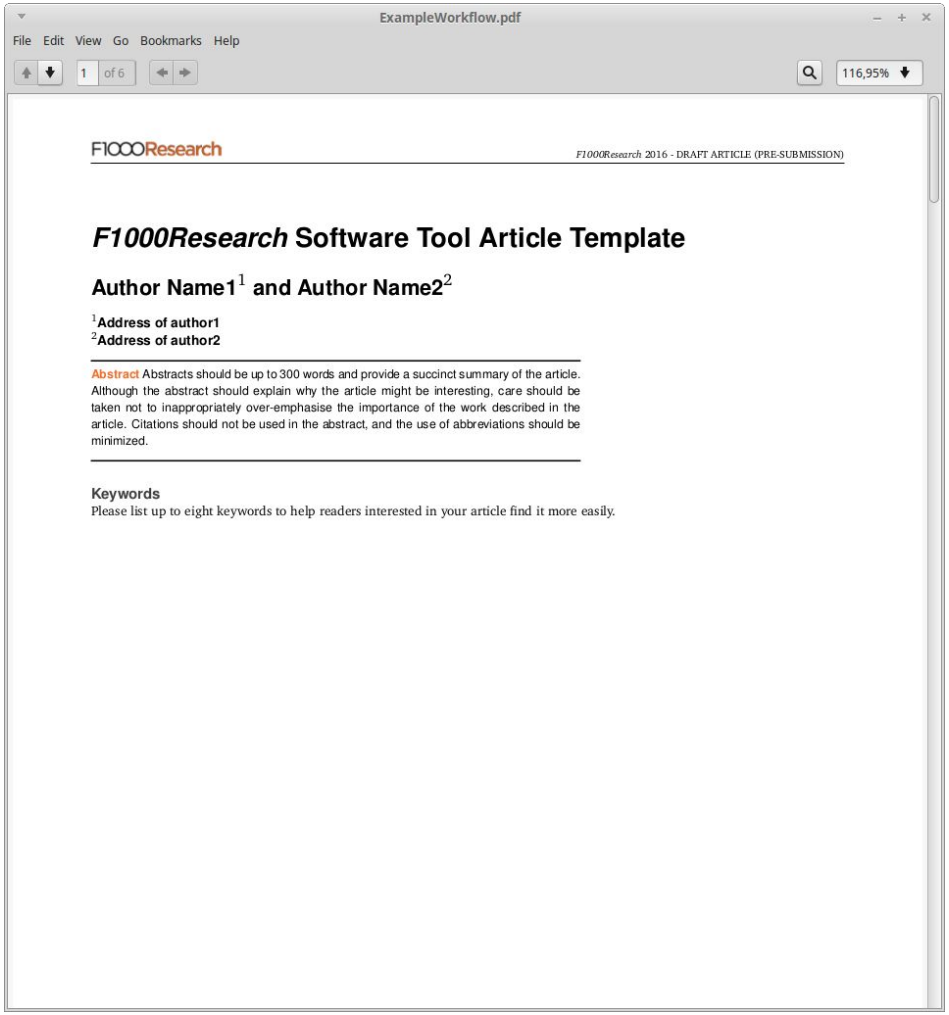
Article Template

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



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 commited 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

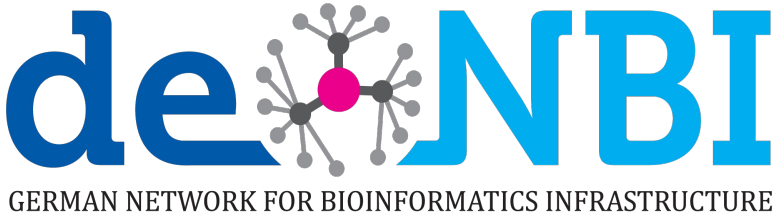
Andrzej Oleś

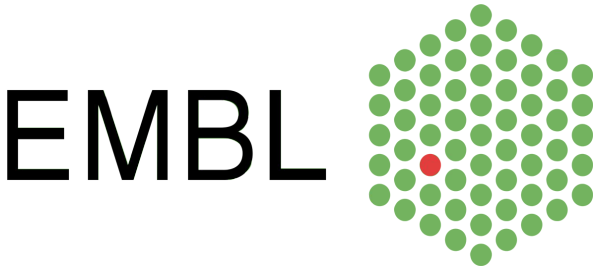
Wolfgang Huber

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