Let's Git With It

Andrew Borgman

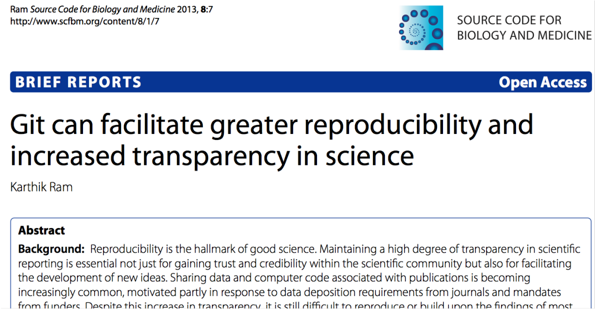
Van Andel Research Institute

September 5, 2013

Overview

1. Git Comfortable
2. Journal Club Business
    1. A few slides covering the basics
3. Xenophobia Musings
    1. What did you think?
    2. Could we do it?
    3. Would it be worth it?
    4. I think so.

Journal Club



Ram *Source Code for Biology and Medicine* 2013, **8**:7
http://www.scfbm.org/content/8/1/7

SOURCE CODE FOR
BIOLOGY AND MEDICINE

BRIEF REPORTS **Open Access**

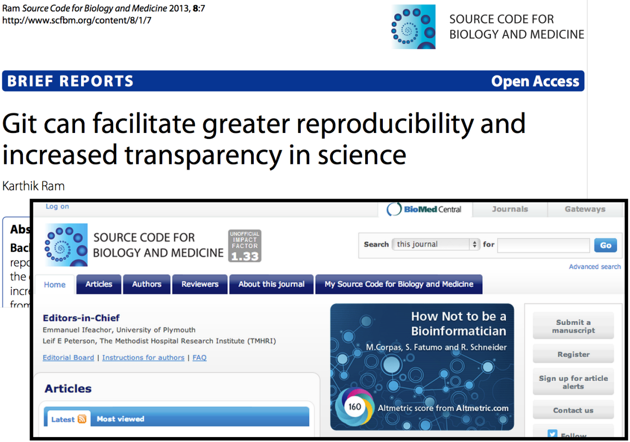
Git can facilitate greater reproducibility and increased transparency in science

Karthik Ram

Abstract

Background: Reproducibility is the hallmark of good science. Maintaining a high degree of transparency in scientific reporting is essential not just for gaining trust and credibility within the scientific community but also for facilitating the development of new ideas. Sharing data and computer code associated with publications is becoming increasingly common, motivated partly in response to data deposition requirements from journals and mandates from funders. Despite this increase in transparency, it is still difficult to reproduce or build upon the findings of most

Journal Club



Ram *Source Code for Biology and Medicine* 2013, **8**:7
http://www.scfbm.org/content/8/1/7

SOURCE CODE FOR BIOLOGY AND MEDICINE

BRIEF REPORTS **Open Access**

Git can facilitate greater reproducibility and increased transparency in science

Karthik Ram

Abs

Bac
repo
the
incr
from

Log on | BioMed Central | Journals | Gateways

SOURCE CODE FOR BIOLOGY AND MEDICINE

UNOFFICIAL IMPACT FACTOR 1.33

Search this journal for Go

Advanced search

Home | Articles | Authors | Reviewers | About this journal | My Source Code for Biology and Medicine

Editors-in-Chief

Emmanuel Ifeachor, University of Plymouth
Leif E Peterson, The Methodist Hospital Research Institute (TMHRI)

Editorial Board | Instructions for authors | FAQ

Articles

Latest | Most viewed

How Not to be a Bioinformatician
M.Corpas, S. Fatumo and R. Schneider

160 Altmetric score from Altmetric.com

Submit a manuscript

Register

Sign up for article alerts

Contact us

Follow

Important Themes

- Open Science Initiatives
  - "Reproducibility is the hallmark of good science."
  - Journals requiring data and analysis code be shared
  - Science needs to be community based to be successful
    - How do I solve a programming problem: Stack Overflow
    - How do I solve a bioinformatics problem: SEQanswers
    - How do I solve a statistics problem: CrossValidated
  - Share your ideas and code!
- Git Can Help Facilitate Open Science
  - Enables reproducibility
  - Provides rigorous tracking of projects
  - GitHub provides community and great hosting
  - BitBucket might be better (https://bitbucket.org/)

Uses Of Git In Science

1. Lab notebook
2. Facilitating collaboration
3. Backup and failsafe against data loss
4. Freedom to explore new ideas and methods
5. Mechanism to solicit feedback and reviews
6. Increase transparency and verifiability
7. Managing large data
8. Lowering barriers to reuse

Why Should We Do It?

- Provides high level of integrity in analysis
  - Results can always be reproduced by typing 'make'
  - All figures and tables created directly from data
  - Everything is tracked; no more uncertainty
- Should be providing "publication quality analysis"
  - Our duty as a core service to VARI investigators
  - Publications are asking for data and code; have it ready
- We need more structure
  - Current structure is OK, but it is not effective for code sharing
  - Still can't step in a project dir and know what is going on
  - I think Git/Make approach would help

How Should We Do It – Logistics

- Up for debate
- Use a public hosting site?
  - Initially thought GitHub would be best
    - Can pay small monthly fee to have private repos
  - BitBucket allows for unlimited private repos
    - Pay for # of users
    - Free for 5, $10/mo for 10
    - This might be the way to go for us
  - BitBucket example: https://bitbucket.org/borgmaan/xenophobia/overview
- Host it ourselves?
  - This might get too hairy
  - Would let us store and track **all** data though

How Should We Do It – Day-to-Day

- Keep it simple to allow for quick integration
- Master standard workflow
  - Create, add, push, pull
- Only need a few commands
  - init, add, commit, push pulll
- Can follow my structure
  - Use R to create all images and tables (TEX) directly from your data
  - Use LYX for a MS Word substitute for report creation
  - Export LATEX source from LYX file
  - Create Makefile that runs your R code, creates your figures, then compiles your report using pdflatex or something like that
  - http://kbroman.github.io/minimal_make/ :: Like this!

Questions?

On to a demo @ https://github.com/borgmaan/xenophobia!