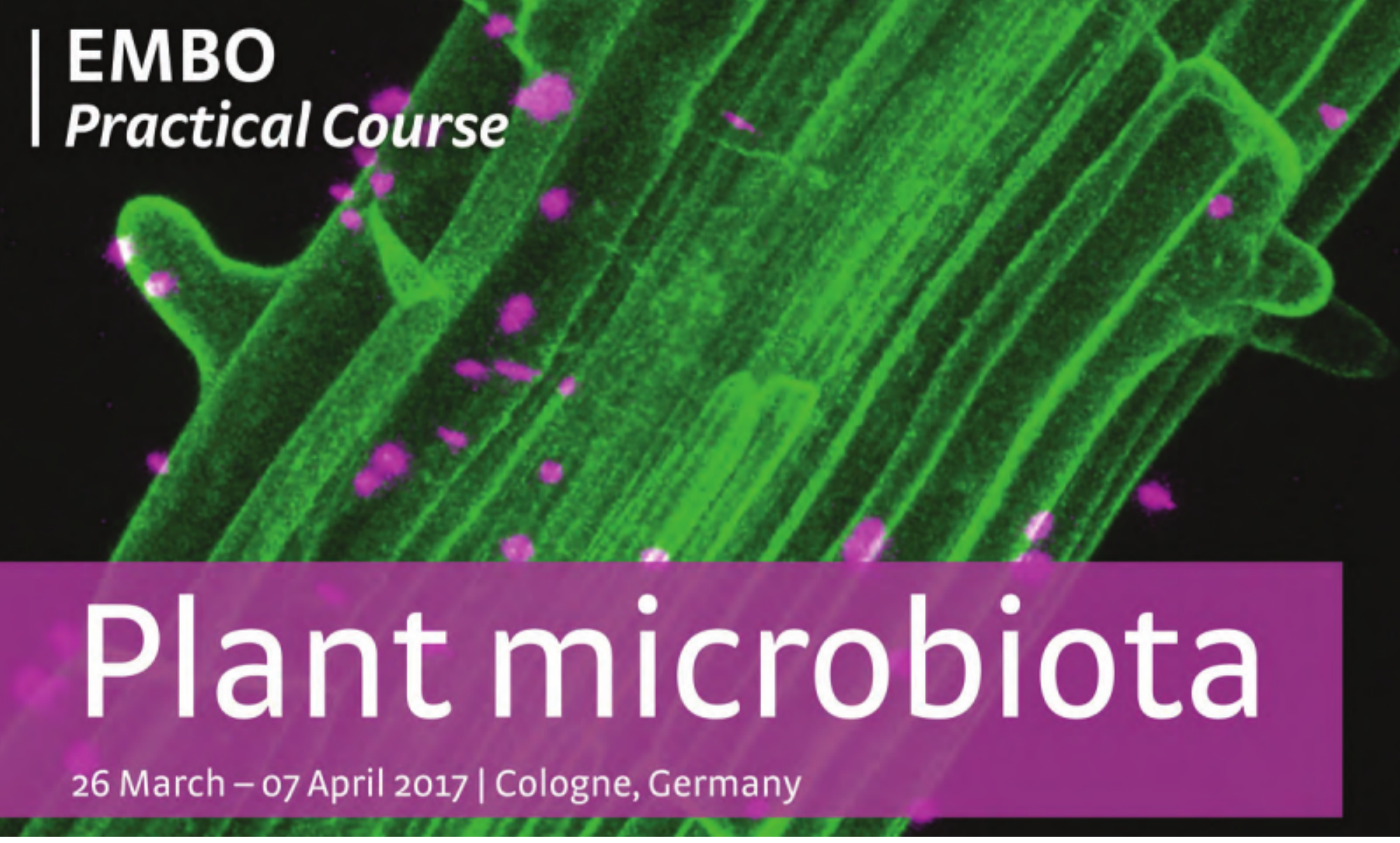


Transparency and reproducibility in computational biology

Ruben Garrido-Oter

Replicability v. reproducibility

reproducibility: someone else using independently generated data and independently developed tools obtaining similar results

replicability: someone else using the same data and the same tools obtaining the same results

in computational biology, reproducibility starts with replicability

Reproducible research in computational biology

it is necesary to provide:

- **raw** data (e.g. straight from the sequencer)
- source code or scripts
- proper documentation (provided with the code)
- detailed method description (normally in the paper)

Now standard in our lab

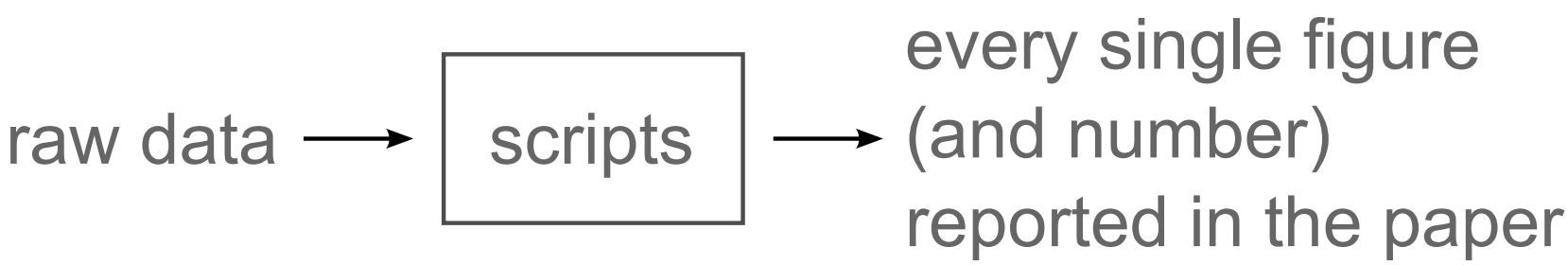
http://www.mpipz.mpg.de/R_scripts

Now standard in our lab

http://www.mpipz.mpg.de/R_scripts

despite our best efforts,
there is a lot of room for improvement!

Ideal situation



The reality

most experimental papers, even in top journals, do not provide ANY code

some provide scripts "upon request"

very few fully (or even partially!) replicable

Ian Holmes
@ianholmes

You can download our code from the URL supplied. Good luck downloading the only postdoc who can get it to run, though

Producing fully reproducible computational research is **hard**

for many steps, it takes much more time to write the code to do them automatically than doing them manually (once or even a few times)

(although in the long run it often pays off)

Partial replicability is better
than no replicability!

Version control tools

necessary to manage the development
of software and code

allow to keep track of changes and versions
for safety and transparency

useful to coordinate projects with several people involved

Stablished version control tools

- GIT (github)
- Subversion (SVN)
- Mercurial
- Concurrent Versions System (CVS)

Stablished version control tools

- **GIT (github)**
- Subversion (SVN)
- Mercurial
- Concurrent Versions System (CVS)

Basic guidelines for good practices in computational biology

- provide all raw, unprocessed data
- deposite scripts and use a version control system
- ideally full replicability (but partial is better than none)
- document code as much as possible

Thank you!