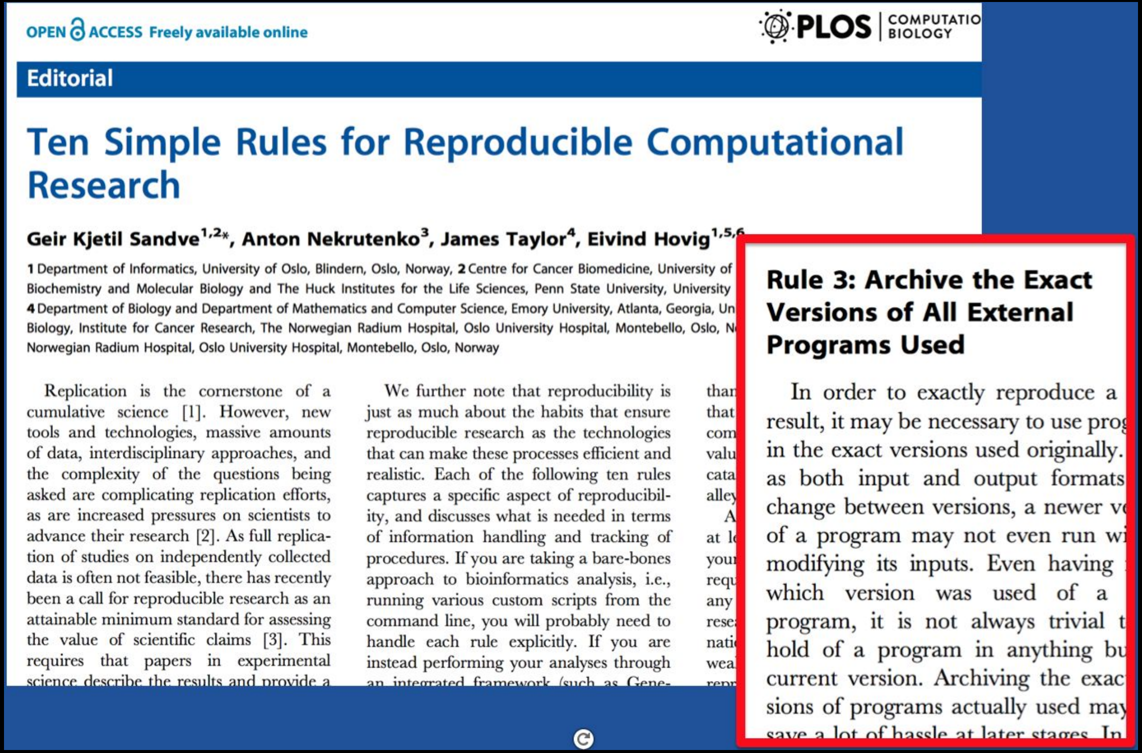
bioaRchive: Improving reproducibility of Bioconductor analyses in Galaxy

Nitesh Turaga, Eric Rasche, Enis Afgan, Dannon Baker, Galaxy Team

July 7-8th 2015
GCC, Norwich 2015

As a community we should be focusing on completely reproducible analysis.



OPEN ACCESS Freely available online

PLOS | COMPUTATIO BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, 2 Centre for Cancer Biomedicine, University of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University 4 Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, Un Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, N Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

Replication is the cornerstone of a cumulative science [1]. However, new tools and technologies, massive amounts of data, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts, as are increased pressures on scientists to advance their research [2]. As full replication of studies on independently collected data is often not feasible, there has recently been a call for reproducible research as an attainable minimum standard for assessing the value of scientific claims [3]. This requires that papers in experimental science describe the results and provide a

We further note that reproducibility is just as much about the habits that ensure reproducible research as the technologies that can make these processes efficient and realistic. Each of the following ten rules captures a specific aspect of reproducibility, and discusses what is needed in terms of information handling and tracking of procedures. If you are taking a bare-bones approach to bioinformatics analysis, i.e., running various custom scripts from the command line, you will probably need to handle each rule explicitly. If you are instead performing your analyses through an integrated framework (such as Gene-

than
that
com
valu
cata
alley
A
at l
your
requ
any
rese
nati
wea
repr

Rule 3: Archive the Exact Versions of All External Programs Used

In order to exactly reproduce a result, it may be necessary to use prog in the exact versions used originally. as both input and output formats change between versions, a newer v of a program may not even run wi modifying its inputs. Even having which version was used of a program, it is not always trivial t hold of a program in anything bu current version. Archiving the exac sions of programs actually used may save a lot of hassle at later stages. In

Bioconductor is a highly used bioinformatics software toolkit.

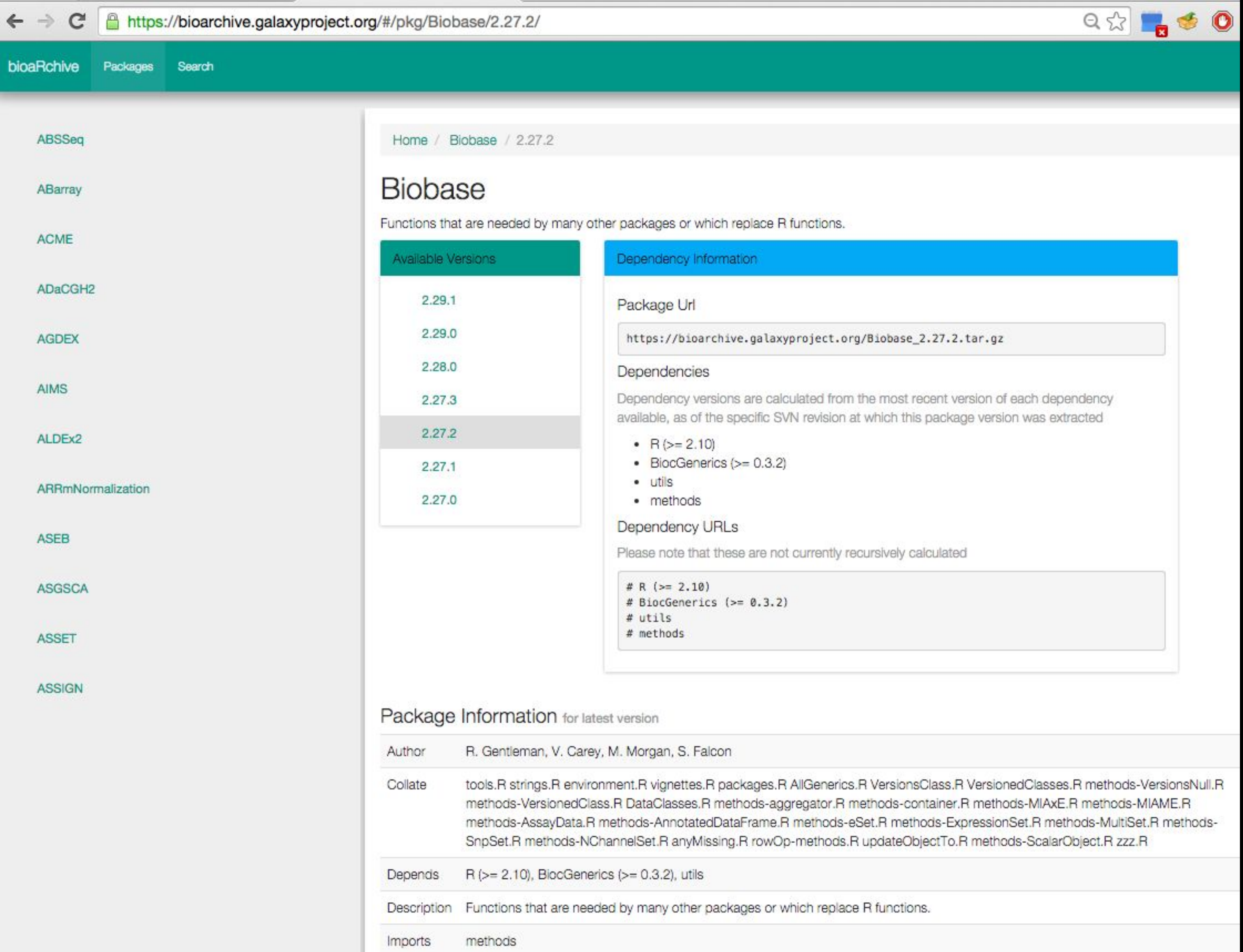
BUT ….

ONLY the most recent version of any tool is available to the user.

What is bioaRchive?

Repository of all versions of all Bioconductor packages and these can be easily obtained from

bioarchive.galaxyproject.org



Biobase

Functions that are needed by many other packages or which replace R functions.

Available Versions

- 2.29.1
- 2.29.0
- 2.28.0
- 2.27.3
- 2.27.2
- 2.27.1
- 2.27.0

Dependency Information

Package Url

```
https://bioarchive.galaxyproject.org/Biobase_2.27.2.tar.gz
```

Dependencies

Dependency versions are calculated from the most recent version of each dependency available, as of the specific SVN revision at which this package version was extracted

- R (>= 2.10)
- BiocGenerics (>= 0.3.2)
- utils
- methods

Dependency URLs

Please note that these are not currently recursively calculated

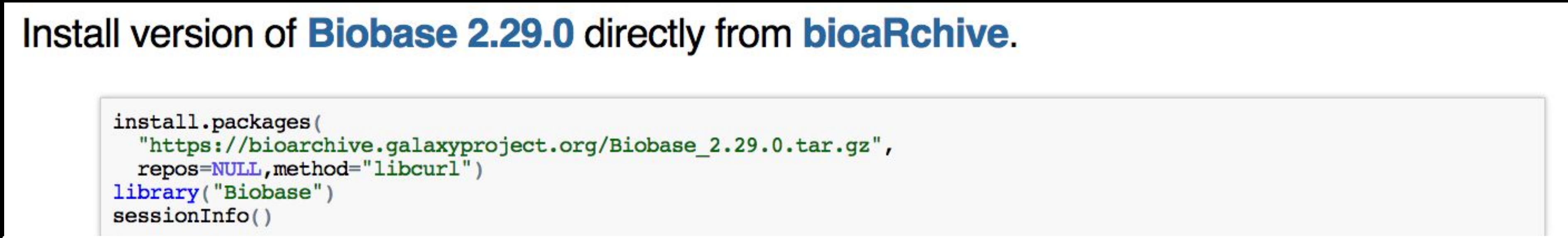
```
# R (>= 2.10)
# BiocGenerics (>= 0.3.2)
# utils
# methods
```

Package Information for latest version

Author	R. Gentleman, V. Carey, M. Morgan, S. Falcon
Collate	tools.R strings.R environment.R vignettes.R packages.R AllGenerics.R VersionsClass.R VersionedClasses.R methods-VersionsNull.R methods-VersionedClass.R DataClasses.R methods-aggregator.R methods-container.R methods-MIAxE.R methods-MIAME.R methods-AssayData.R methods-AnnotatedDataFrame.R methods-eSet.R methods-ExpressionSet.R methods-MultiSet.R methods-SnpSet.R methods-NChannelSet.R anyMissing.R rowOp-methods.R updateObjectTo.R methods-ScalarObject.R zzz.R
Depends	R (>= 2.10), BiocGenerics (>= 0.3.2), utils
Description	Functions that are needed by many other packages or which replace R functions.
Imports	methods

Thanks to Eric Rasche for the UI.

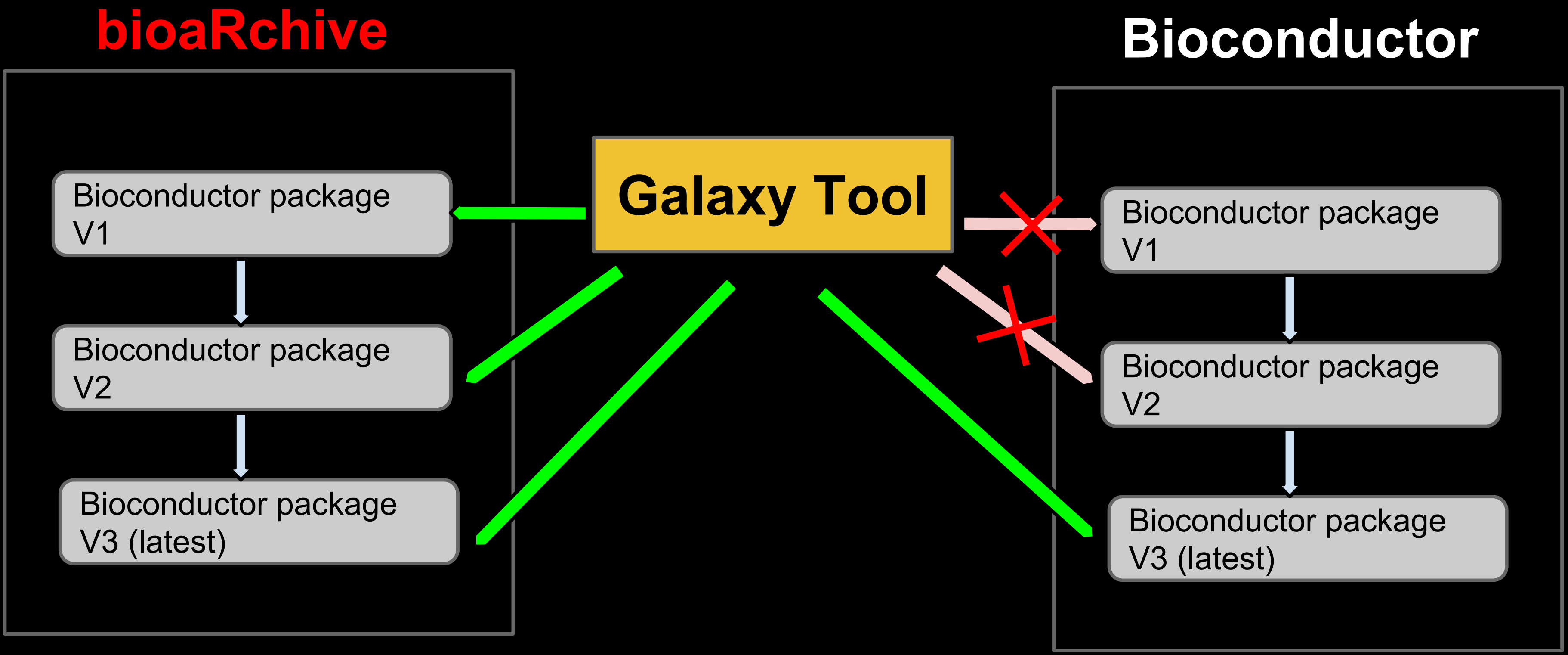
Using bioaRchive



Install version of Biobase 2.29.0 directly from bioaRchive.

```
install.packages(
  "https://bioarchive.galaxyproject.org/Biobase_2.29.0.tar.gz",
  repos=NULL,method="libcurl")
library("Biobase")
sessionInfo()
```

Also facilitating Bioconductor and Galaxy interoperability



bioaRchive

Bioconductor package V1

Bioconductor package V2

Bioconductor package V3 (latest)

Galaxy Tool

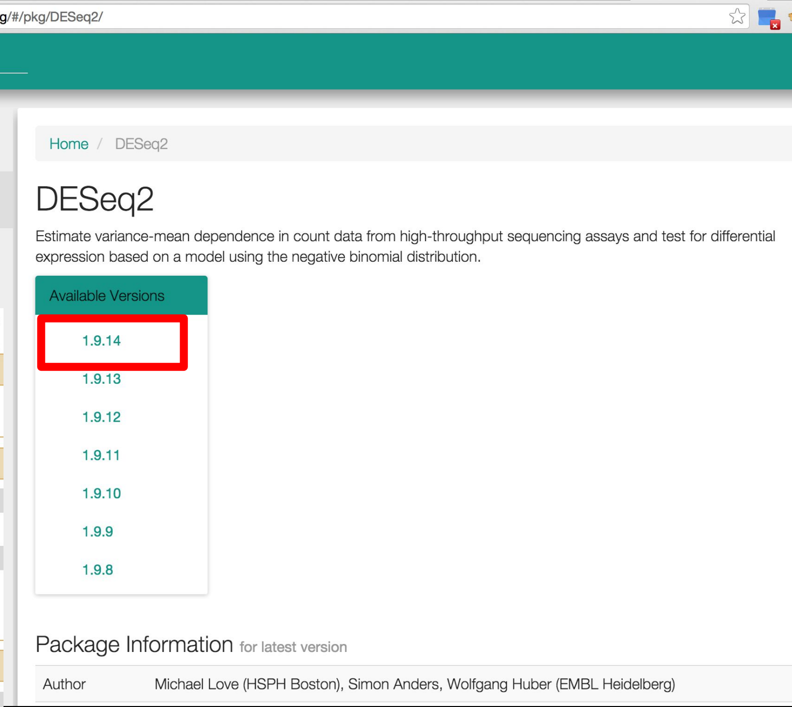
Bioconductor

Bioconductor package V1

Bioconductor package V2

Bioconductor package V3 (latest)

Ooops!



DESeq2

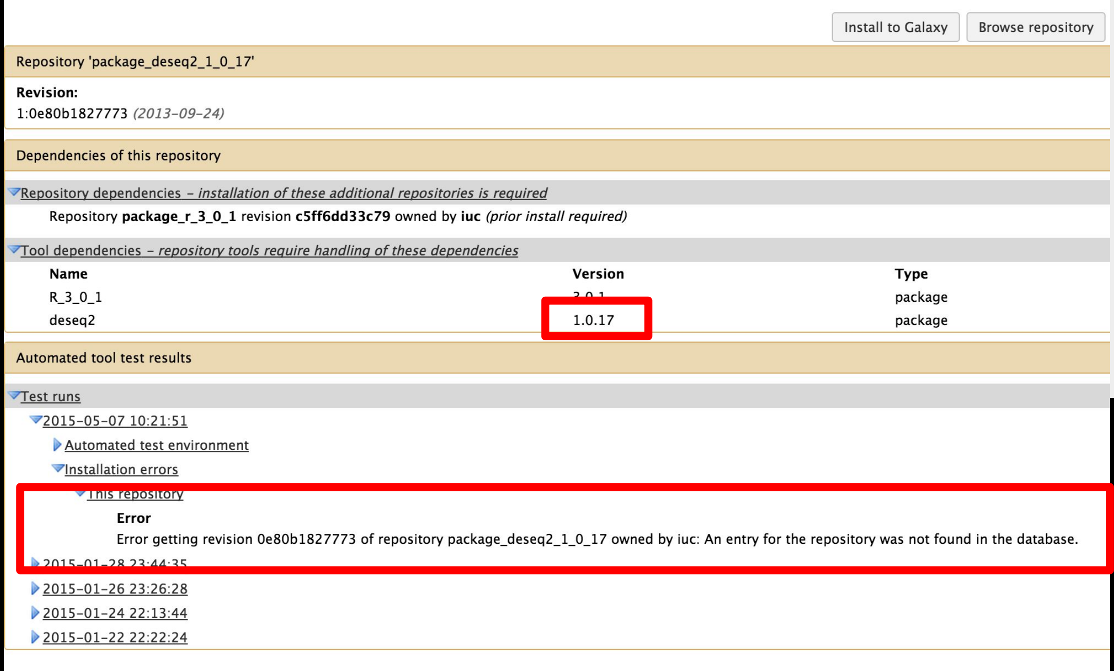
Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Available Versions

- 1.9.14
- 1.9.13
- 1.9.12
- 1.9.11
- 1.9.10
- 1.9.9
- 1.9.8

Package Information for latest version

Author	Michael Love (HSPH Boston), Simon Anders, Wolfgang Huber (EMBL Heidelberg)



Repository 'package_deseq2_1_0_17'

Revision:
1:0e80b1827773 *(2013-09-24)*

Dependencies of this repository

Repository dependencies – *installation of these additional repositories is required*

Repository **package_r_3_0_1** revision **c5ff6dd33c79** owned by **iuc** *(prior install required)*

Tool dependencies – *repository tools require handling of these dependencies*

Name	Version	Type
R_3_0_1	3.0.1	package
deseq2	1.0.17	package

Automated tool test results

Test runs

- 2015-05-07 10:21:51
  - Automated test environment
  - Installation errors
    - This repository

Error

Error getting revision 0e80b1827773 of repository package_deseq2_1_0_17 owned by iuc: An entry for the repository was not found in the database.

- 2015-01-28 23:44:35
- 2015-01-26 23:26:28
- 2015-01-24 22:13:44
- 2015-01-22 22:22:24

Resolve dependencies from bioarchive directly for galaxy tools.

Future work

Improve dependency management for bioconductor based analysis

- Missing versions of dependencies
- Packages with multiple dependencies.

Want to help?

bioarchive.github.io

Contact:

nitesh.turaga@gmail.com