Legal *Inter*-operability & *Intra*-operability of Research Data:

The Case of the Research Compendium

Gail Clement,
RDA/CODATA Legal Interop IG Co-Chair
Head of Research Services, Caltech Library
https://orcid.org/0000-0001-5494-4806 | @ repositorian

So, what is a Research Compendium?

1.3 Research compendia

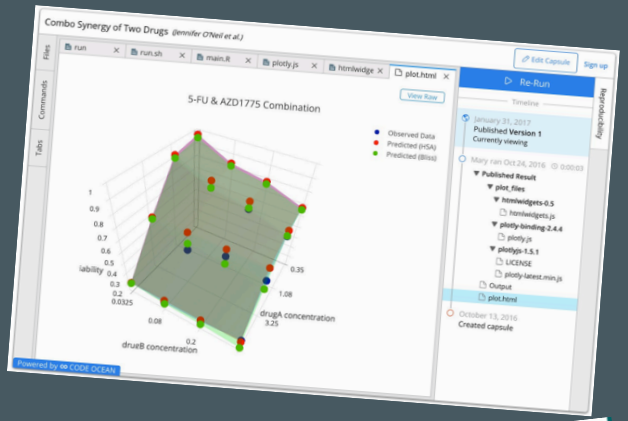
The term *research compendium* was coined by Gentleman and Lang (2007) who "introduce[d] the concept of a compendium as both a container for the different elements that make up the document and its computations (i.e. text, code, data,...), and as a means for distributing, managing and updating the collection." According to Marwick, Boettiger, and Mullen (2018) it provides "a standard and easily recognisable way for organising the digital materials of a research project to enable other researchers to inspect, reproduce, and extend the research". This standard may differ between scien-

Daniel Nüst, Carl Boettiger, Ben Marwick, *How to Read a Research Compendium*, arXiv:1806.09525v1 [cs.GL] 11Jun2018

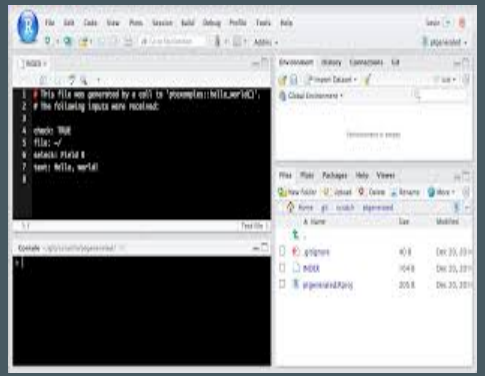
Also known as ….

Jupyter notebooks

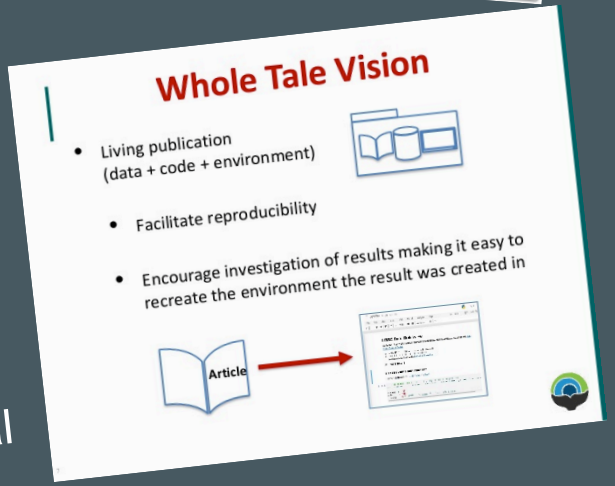
Compute capsules



Interactive Notebooks

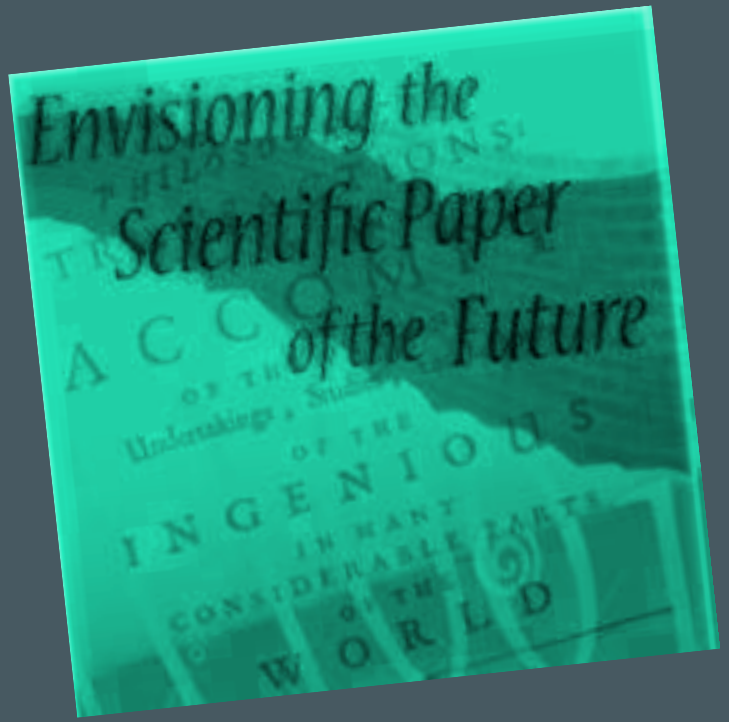


R Projects
R Notebooks



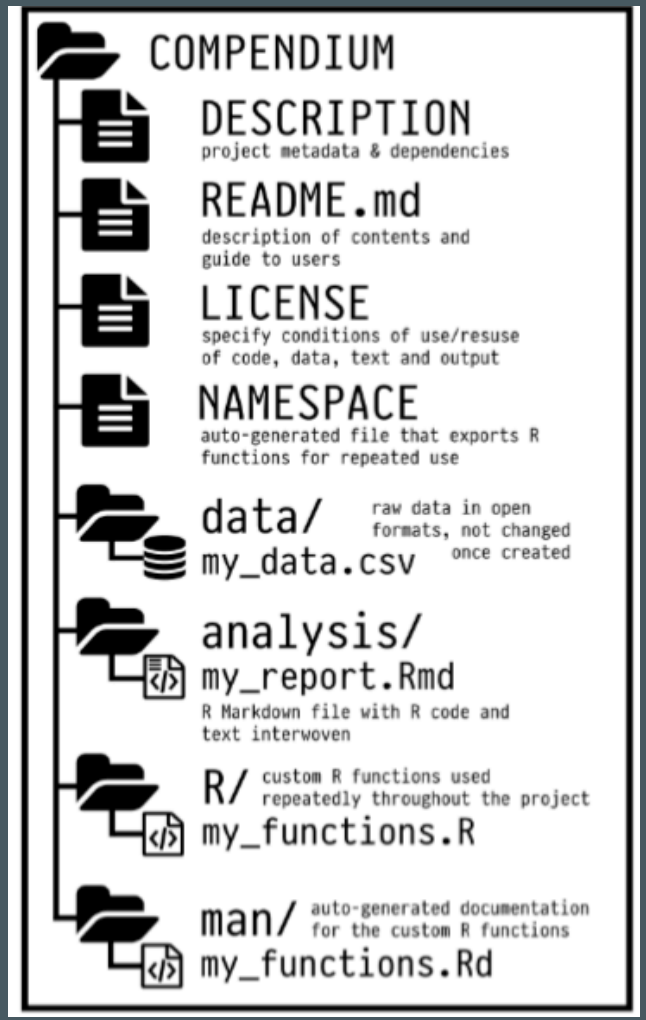
Computational Narratives

Also known as ....



Living Paper

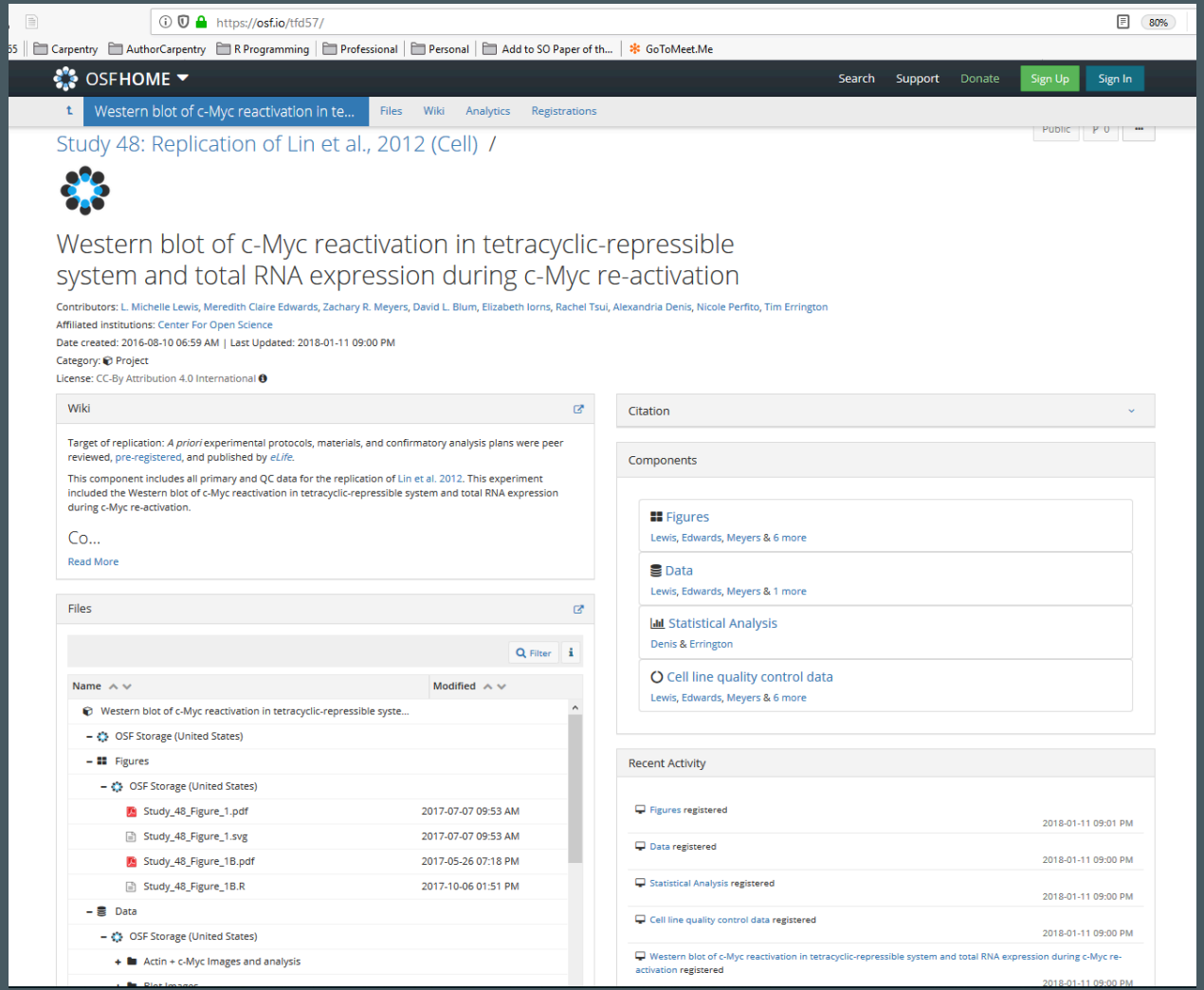
eLife's new computationally reproducible article will allow users to modify figures



Characteristics of the Research Compendium

- Data rich digital research object
- Combines text and markup, underlying data, code for analysis , figures/plots
- Multiple files representing diverse formats and types of digital objects
- Electronically packaged into a compound object for digital distribution

Marwick, Ben, Carl Boettiger, and Lincoln Mullen. “Packaging Data Analytical Work Reproducibly Using R (and Friends).” The American Statistician 72 (1):80–88.https://doi.org/10.1080/00031305.2017.1375986.



Characteristics of the Research Compendium

Why worry about the Legal Interoperability of a Research Compendium ?

Why worry about licensing of Research Compendium

- Gaining wider interest among research authors Early adopter publishers are attracting buzz with their experiments
- Observable confusion and misunderstanding
- Hard to find good examples of licensing done right
- Not-so-hard to find examples of worrisome licensing practices

Licensing for research compendia?

Best Practices reproducibility, research-compendia

As I was creating the repo and was confronted with assigning a license, I was stumped on how best to license this project. In @benmarwick, @cboettig, and @lincoln paper the discuss licensing and follow the suggestions in https://web.stanford.edu/~vcs/papers/ERROLSI03092009.pdf. Essentially the different parts of a research compendia should have different licenses. For example, the manuscript and figures could use CC-BY, the code MIT, and the data CC0.

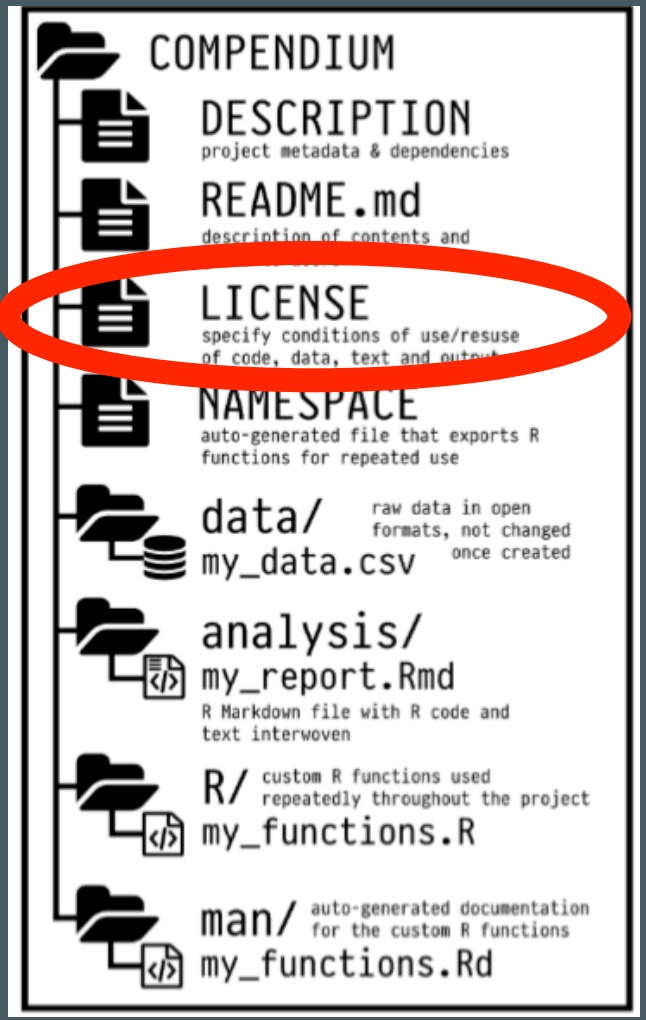
My question is more on the nuts and bolts best practice of capturing the fact that a compendia would be released under the multiple licenses. Should I:

- List License: CC-BY, MIT, CC0 in the DESCRIPTION
- Explain in a README
- Both
- Something else?

I've dug around and can't find an implementation of this multiple license concept for compendia. Thoughts most welcome!

https://discuss.ropensci.org/t/licensing-for-research-compendia/1581

Question 1.
Won’t one omnibus license cover the whole aggregation of objects?



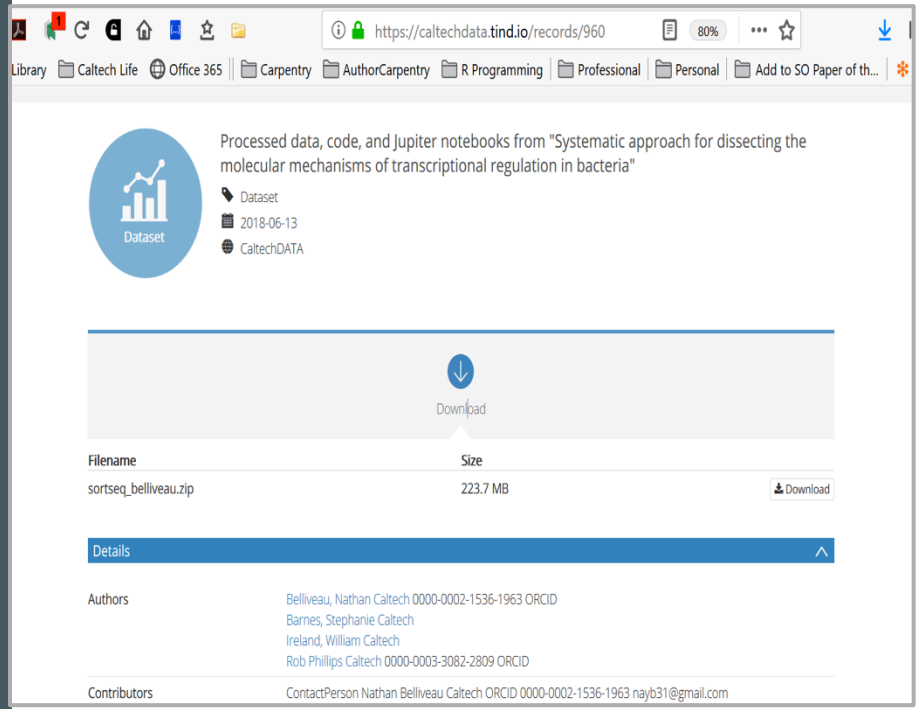
LICENSE

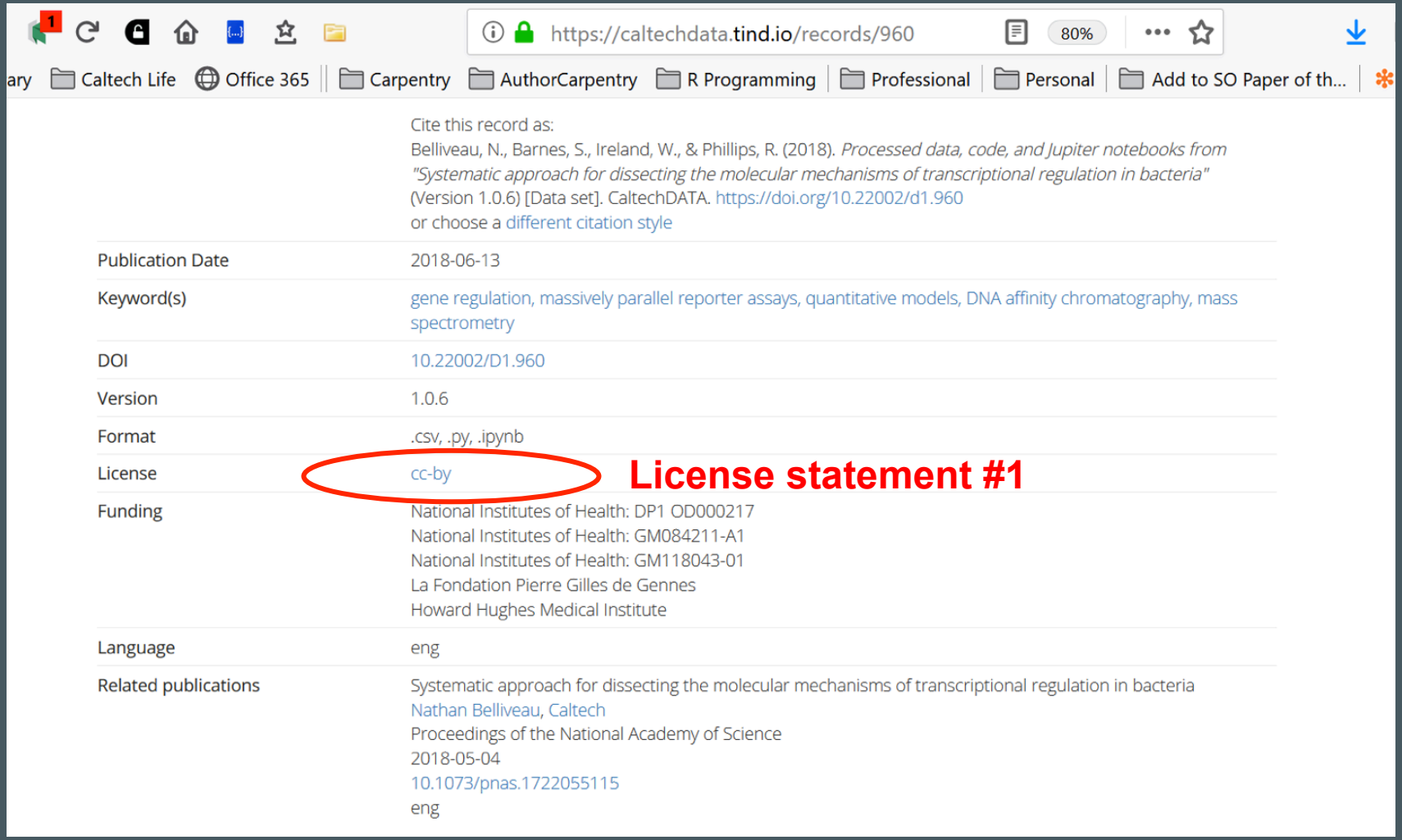
specify conditions of use/resuse
of code, data, text and output

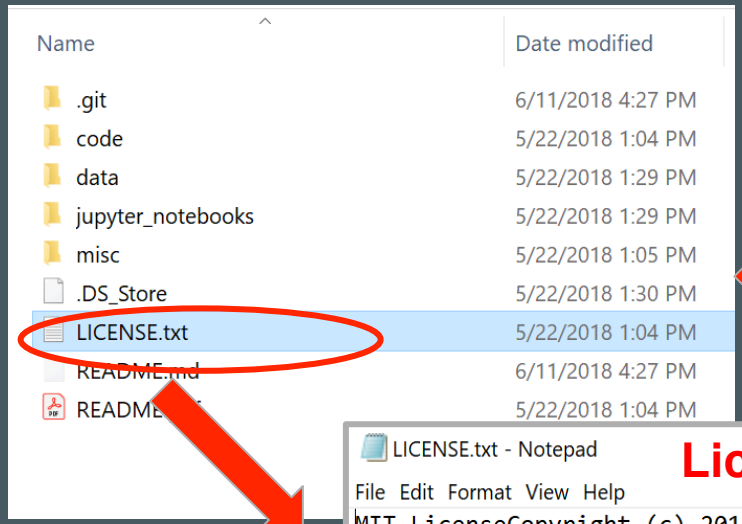
Marwick, Ben, Carl Boettiger, and Lincoln Mullen. “Packaging Data Analytical Work Reproducibly Using R (and Friends).” The American Statistician 72 (1): 80–88.https://doi.org/10.1080/00031305.2017.1375986.

Example 1:

Jupyter Notebook (and related files) zipped as one record in CaltechData Repository and used locally with relevant installed dependencies




https://caltechdata.tind.io/records/960
80%
Caltech Life
Office 365
Carpentry
AuthorCarpentry
R Programming
Professional
Personal
Add to SO Paper of th...
Cite this record as:
Belliveau, N., Barnes, S., Ireland, W., & Phillips, R. (2018). Processed data, code, and Jupiter notebooks from "Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria" (Version 1.0.6) [Data set]. CaltechDATA. https://doi.org/10.22002/d1.960
or choose a different citation style
Publication Date
2018-06-13
Keyword(s)
gene regulation, massively parallel reporter assays, quantitative models, DNA affinity chromatography, mass spectrometry
DOI
10.22002/D1.960
Version
1.0.6
Format
.csv, .py, .ipynb
License
cc-by
License statement #1
Funding
National Institutes of Health: DP1 OD000217
National Institutes of Health: GM084211-A1
National Institutes of Health: GM118043-01
La Fondation Pierre Gilles de Gennes
Howard Hughes Medical Institute
Language
eng
Related publications
Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria
Nathan Belliveau, Caltech
Proceedings of the National Academy of Science
2018-05-04
10.1073/pnas.1722055115
eng



The downloaded and unzipped files look like this on the desktop

LICENSE.txt - Notepad

License statement #2

File Edit Format View Help

MIT LicenseCopyright (c) 2017 Rob Phillips group @ California Institute of TechnologyPermission is hereby granted, free of charge, to any person obtaining a copyof this software and associated documentation files (the "Software"), to dealin the Software without restriction, including without limitation the rightsto use, copy, modify, merge, publish, distribute, sublicense, and/or sellcopies of the Software, and to permit persons to whom the Software isfurnished to do so, subject to the following conditions:The above copyright notice and this permission notice shall be included in allcopies or substantial portions of the Software. THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS ORIMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THEAUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHERLIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THESOFTWARE.

jupyter SI_B_mutation_diversity Last Checkpoint: 05/22/2018 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3

Markdown

Supplemental Information Section B - Analysis of library diversity

In [1]:
```
import scipy.ndimage

# Our numerical workhorses
import numpy as np
import pandas as pd
import scipy as sp

# Import the project utils
import sys
sys.path.insert(0, '../code/')
import NB_sortseq_utils as utils

# Import matplotlib stuff for plotting
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from IPython.core.pylabtools import figsize

# Seaborn, useful for graphics
```

License statement #3

jupyter SI_H_thermodynamic_model_MCMC_processing Last Checkpoint: 10 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Code

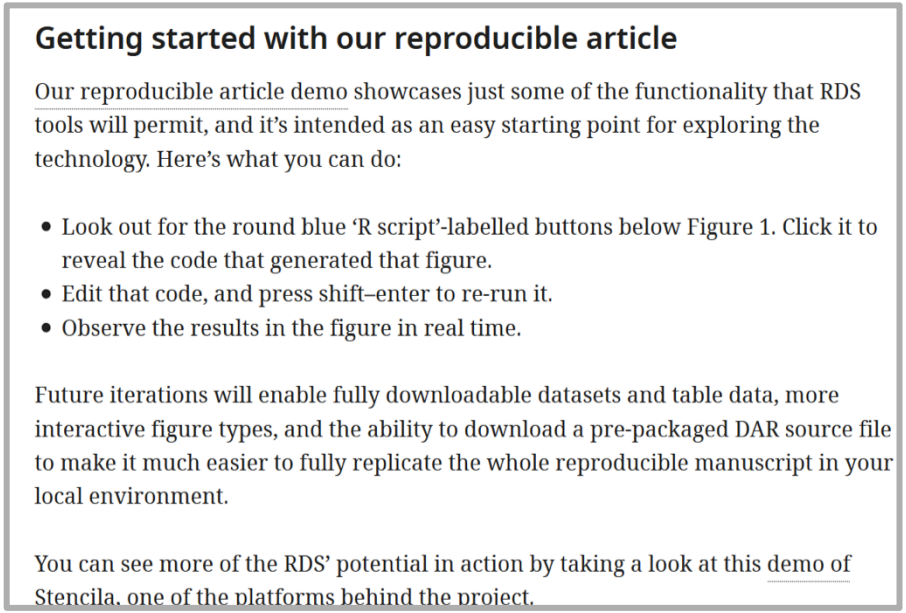
```
the promoter as having a simple repression architecture (Bintu
Some additional complexity arises due to the presence of other
sites on the genome, and the allosteric dependence of a purine
co-repression. Following the approach of  Weinert _et al._
(Weinert _et al._, 2014), this can be quantitatively describe

\begin{equation}
P_{bound} = \frac{\lambda_p e^{-\beta \varepsilon_p}} {1 +
\lambda_p e^{-\beta \varepsilon_p} + \lambda_r e^{-\beta \va
\end{equation}
Here $\lambda_p$ and $\lambda_r$ represent the fugacity, whi
relative availability of RNAP and PurR, respectively, to bi
sites. These parameters depend on the concentration of each
their chemical potentials), and for PurR, will also depend
state. $\varepsilon_p$ and $\varepsilon_r$ represent the binding energies of
RNAP and PurR to their binding sites, respectively.
```

In a computational narrative, which part is the "text" (cc-by license) and which part is the "code"? (MIT license)?

A machine may not see the difference!

Example 2:

Reproducible article published in Elife with a Binder-ized manifestion of the article to enable interactivity online by anyone



Getting started with our reproducible article

Our reproducible article demo showcases just some of the functionality that RDS tools will permit, and it's intended as an easy starting point for exploring the technology. Here's what you can do:

- Look out for the round blue 'R script'-labelled buttons below Figure 1. Click it to reveal the code that generated that figure.
- Edit that code, and press shift–enter to re-run it.
- Observe the results in the figure in real time.

Future iterations will enable fully downloadable datasets and table data, more interactive figure types, and the ability to download a pre-packaged DAR source file to make it much easier to fully replicate the whole reproducible manuscript in your local environment.

You can see more of the RDS' potential in action by taking a look at this demo of Stencila, one of the platforms behind the project.

https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article

CANCER BIOLOGY, BIOCHEMISTRY AND CHEMICAL BIOLOGY

Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

L Michelle Lewis, Meredith C Edwards, Zachary R Meyers, C Conover Talbot Jr, Haiping Hao, David Blum, Reproducibility Project: Cancer Biology
University of Georgia, Bioexpression and Fermentation Facility, United States; Johns Hopkins University, Deep Sequencing and Microarray Core Facility, United States

REPLICATION STUDY Jan 9, 2018

CITED 4 VIEWS 2,794 ANNOTATIONS 2

CITE AS: eLife 2018;7:e30274 DOI: 10.7554/eLife.30274

Abstract

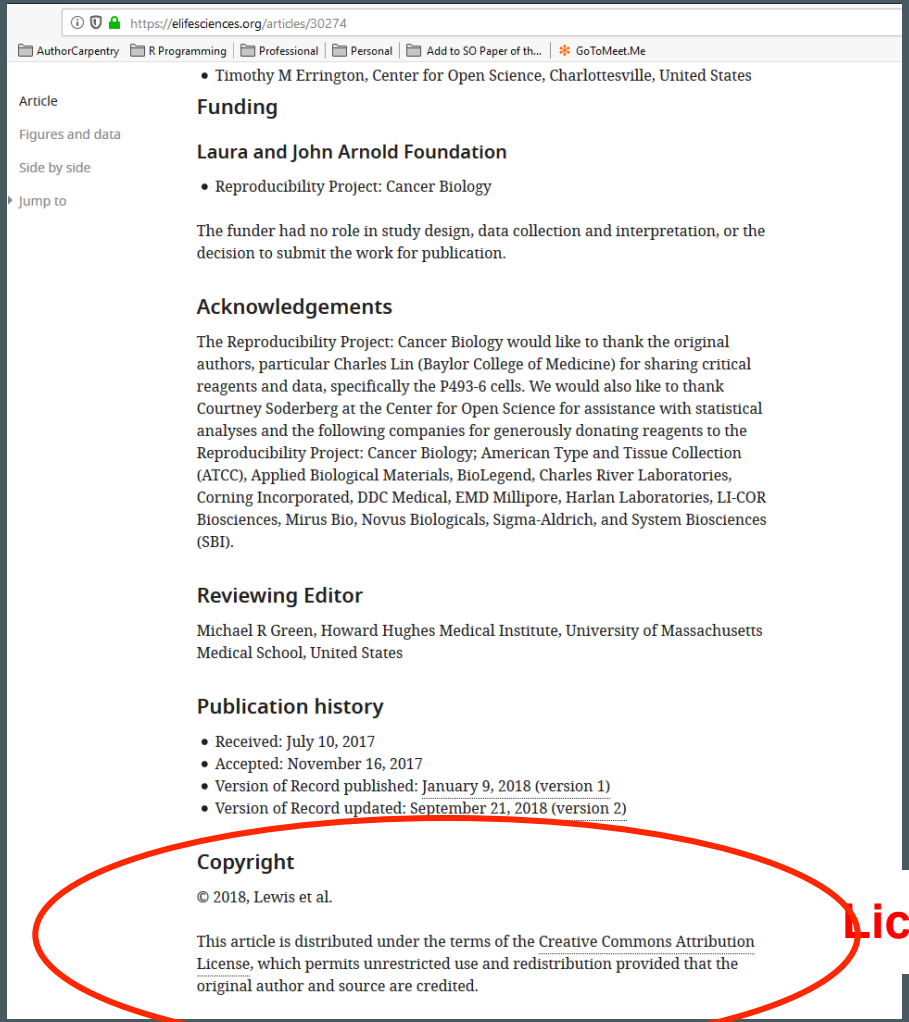
As part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Blum et al., 2015), that described how we intended to replicate selected experiments from the paper 'Transcriptional amplification in tumor cells with elevated c-Myc' (Lin et al., 2012). Here we report the results. We found overexpression of c-Myc increased total levels of RNA in P493-6 Burkitt's lymphoma cells; however, while the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication. Digital gene expression analysis for a set of genes was also performed on P493-6 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression, similar to the original study (Figure 3F; Lin et al., 2012). Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012). Treating the data as paired, we found a statistically significant increase in gene expression

RELATED TO

Registered report: Transcriptional amplification in tumor cells with elevated c-Myc
David Blum et al.

REGISTERED REPORT Jan 26, 2015

Further reading »



https://elifesciences.org/articles/30274

AuthorCarpentry | R Programming | Professional | Personal | Add to SO Paper of th... | GoToMeet.Me

Article

Figures and data

Side by side

Jump to

- Timothy M Errington, Center for Open Science, Charlottesville, United States

Funding

Laura and John Arnold Foundation

- Reproducibility Project: Cancer Biology

The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgements

The Reproducibility Project: Cancer Biology would like to thank the original authors, particular Charles Lin (Baylor College of Medicine) for sharing critical reagents and data, specifically the P493-6 cells. We would also like to thank Courtney Soderberg at the Center for Open Science for assistance with statistical analyses and the following companies for generously donating reagents to the Reproducibility Project: Cancer Biology; American Type and Tissue Collection (ATCC), Applied Biological Materials, BioLegend, Charles River Laboratories, Corning Incorporated, DDC Medical, EMD Millipore, Harlan Laboratories, LI-COR Biosciences, Mirus Bio, Novus Biologicals, Sigma-Aldrich, and System Biosciences (SBI).

Reviewing Editor

Michael R Green, Howard Hughes Medical Institute, University of Massachusetts Medical School, United States

Publication history

- Received: July 10, 2017
- Accepted: November 16, 2017
- Version of Record published: January 9, 2018 (version 1)
- Version of Record updated: September 21, 2018 (version 2)

Copyright

License statement #1

https://repro.elifesciences.org/example.html# 50%

Office 365 | Carpentry | AuthorCarpentry | R Programming | Professional | Personal | Ac

eLIFE ABOUT COMMUNITY SUBMIT MY RESEARCH LOG IN/REGISTER

HOME MAGAZINE INNOVATION

This is a Reproducible document. See the original article or source.

Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

L Michelle Lewis, Meredith C Edwards, Zachary R Meyers, C Conover Talbot Jr, Haiping Hao, David Blum, Reproducibility Project: Cancer Biology

As part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Blum et al., 2015), that described how we intended to replicate selected experiments from the paper 'Transcriptional amplification in tumor cells with elevated c-Myc' (Lin et al., 2012). Here we report the results. We found overexpression of c-Myc increased total levels of RNA in P493-6 Burkitt's lymphoma cells; however, while the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication. Digital gene expression analysis for a set of genes was also performed on P493-6 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression, similar to the original study (Figure 3F; Lin et al., 2012). Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012). Treating the data as paired, we found a statistically significant increase in gene expression for both active and silent genes upon c-Myc induction, with the change in gene expression greater for active genes compared to silent genes. Finally, we report meta-

https://repro.elifesciences.org/example.html#

uthorCarpentry | R Programming | Professional | Personal | Add to SO Paper of th... | GoToMeet.Me

eLIFE ABOUT COMMUNITY SUBMIT MY RESEARCH LOG IN/REGISTER

HOME MAGAZINE INNOVATION

This is a Reproducible document. See the original article or source.

Introduction

Results and discussion

Conditional expression of c-Myc in the B-cell line P493-6

Total RNA levels following c-Myc overexpression

Digital gene expression following c-Myc overexpression

Meta-analyses of original and replicated effects

reasons discussed above and as prespecified in the Registered Report (Blum et al., 2015). For active genes, expression at 0 hr to 1 hr, 0 hr to 24 hr, and 1 hr to 24 hr the meta-analyses were statistically significant ($p=1.12\times10^{-7}$, $p=7.01\times10^{-4}$, $p=0.0129$, respectively). In all comparisons the results were consistent when considering the direction of the effect; however the effect size point estimate of each study (original, replication serum lot one, replication serum lot two) was not within the confidence interval of the other studies. Further, the large confidence intervals of the meta-analysis along with statistically significant Cochran's Q tests suggest heterogeneity between the original and replication studies. For silent genes, the meta-analysis was not statistically significant for gene expression at 0 hr to 1 hr and 1 hr to 24 hr ($p=0.203$, $p=0.0571$, respectively) and the effect size point estimate of each study was not within the confidence interval of the other studies. Similar to the active gene comparisons, the large confidence intervals of the meta-analysis along with statistically significant Cochran's Q tests suggest heterogeneity between the studies. Furthermore, for the 0 hr to 1 hr comparison the original study and replication studies were in opposite directions, while the 1 hr to 24 hr comparison was consistent. Finally, the comparison between 0 hr and 24 hr for silent genes was consistent when considering direction of the effect with a statistically significant meta-analysis ($p=7.10\times10^{-17}$). The point estimate of the original study was not within the confidence intervals of the replication studies; however both replication studies with different serum lots were within the confidence intervals of the original study and each other. Overall, the gene expression analysis indicates that the effect sizes observed from the two serum lots tested in this replication attempt, although not identical, were more similar to each other than to the original study.

This direct replication provides an opportunity to understand the present evidence of these effects. Any known differences, including reagents and protocol differences, were identified prior to conducting the experimental work and described in the Registered Report (Blum et al., 2015). However, this is limited to what was obtainable from the original paper and through communication with the original authors, which means there might be particular features of the original experimental protocol that could be critical, but unidentified. So while some aspects, such as the cell line, induction time course, and the method used to measure gene expression were maintained, others were changed at the time of study design (Blum et al., 2015) that could affect results, such as the analytical approach (Silberzahn et al., 2017) and serum lot (Leek et al., 2010). Furthermore, other aspects were unknown or not easily controlled for. These include variables such as cell line genetic drift (Hughes et al., 2007; Kleensang et al., 2016) or changes in cellular volume that can impact overall transcript abundance (Padovan-Merhar et al., 2015). Whether these or other factors influence the outcomes of this study is open to hypothesizing and further investigation, which is facilitated by direct replications and transparent reporting.

This is a Reproducible document. See the original article or source.

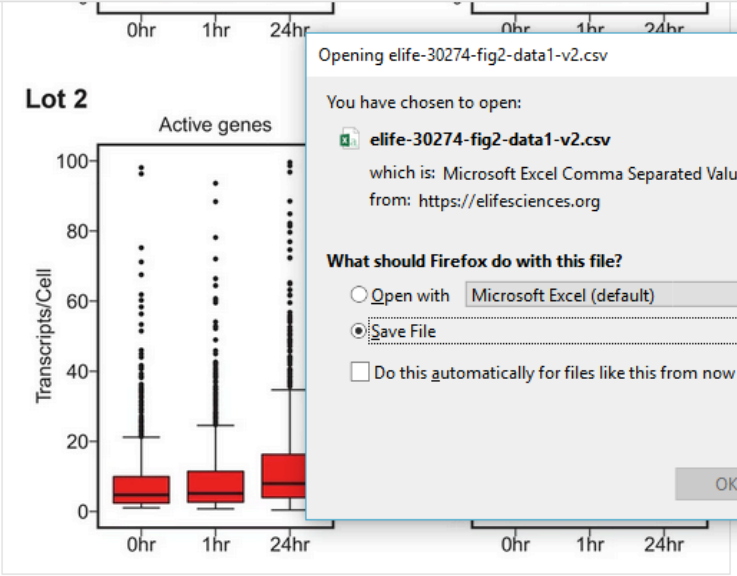
NOTE: Below is a reproducible version of Figure 1B. You can inspect the code, make changes and run the code by pressing SHIFT+ENTER. The data used can be downloaded here.

R Script status: ready (run code with ⇧⏎)

```
library(Rmisc)
library(ggplot2)
library(cowplot)

#names raw data from protocol 2 from csv file
data2 <- read.csv("article/Study_48_Protocol_2_Data.csv", header=T, sep=",")

#creates new column calculating RNA in 100uL
data2$RNA.100uL <- data2$Average.RNA.Concentration*100

##calculates RNA per cell
data2$RNA.per.cell <- data2$RNA.100uL/data2$Total.Cells.Harvested

#calculates RNA per 1000 cells
data2$value <- data2$RNA.per.cell*1000
```

https://elifesciences.org/articles/30274#fig2

AuthorCarpentry | R Programming | Professional | Personal | Add to SO Paper of th... | GoToMeet.Me



Digital gene expression analysis.

P493-6 cells grown in the presence of tetracycline (Tet) for 72 hr for repression of the conditional p*myc*-tet construct, were switched into Tet-free growth medium to induce c-Myc expression. Cells ...

see more »

https://doi.org/10.7554/eLife.30274.003

Figure 2—source data 1

List of Reporter CodeSets and gene expression values.

https://doi.org/10.7554/eLife.30274.006

Download elife-30274-fig2-data1-v2.csv

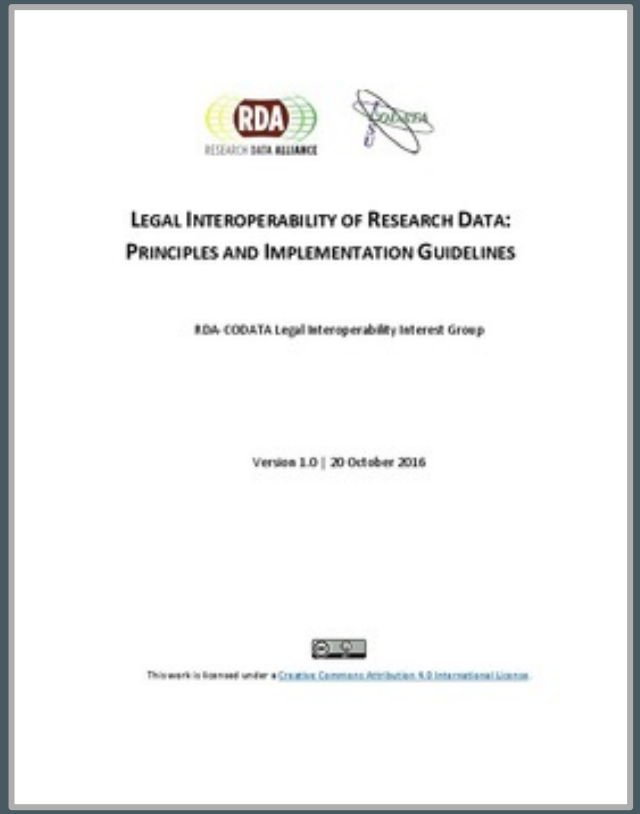
To test whether active genes, as well as silent genes, increased expression during c-Myc induction we performed the confirmatory analysis as outlined in the

Questions about Licensing Practices for the RC (1)

- Does a license in the metadata record inherit all the way through the file directory to apply to all files?
- Can we devise a top level rights statement in the metadata record that links to a more detailed list of licenses associated with each file?
- What is the legal status of a compendium with conflicting licenses?

Questions about Licensing Practices for the RC (2)

- Does the license need to appear in each manifestation of the paper? (HTML, PDF, and the interactive version)
- What does ‘code’ mean in the context of computational narrative where the text itself may be computed?
- Other ????



LEGAL INTEROPERABILITY OF RESEARCH DATA:
PRINCIPLES AND IMPLEMENTATION GUIDELINES

RDA-CODATA Legal Interoperability Interest Group

Version 1.0 | 20 October 2016

This work is licensed under a Creative Commons Attribution 4.0 International License

Please join our IG to discuss many other legal questions around FAIR licensing of Research Compendia!

https://rd-alliance.org/groups/rdacodata-legal-interoperability-ig.html