SevenBridges

Persistent Reproducible Reporting

Nan Xiao, Seven Bridges

2017/05/20 @ China R Conference Beijing

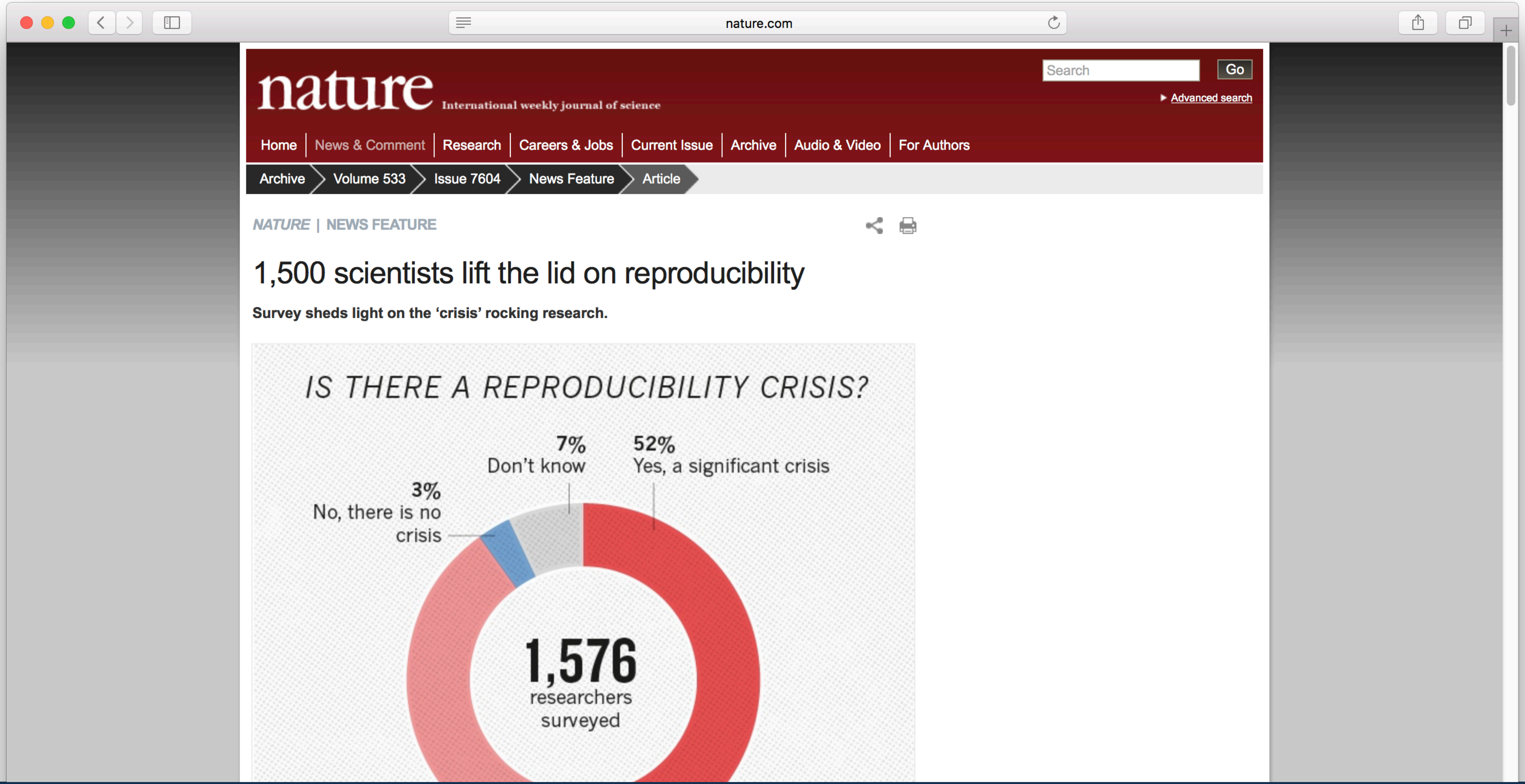
DOCUMENT-LEVEL REPRODUCIBILITY

R MARKDOWN + KNITR TO THE RESCUE

+

REPRODUCIBILITY

… has always been a concern in both academia & industry.



nature.com

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 533 > Issue 7604 > News Feature > Article

NATURE | NEWS FEATURE

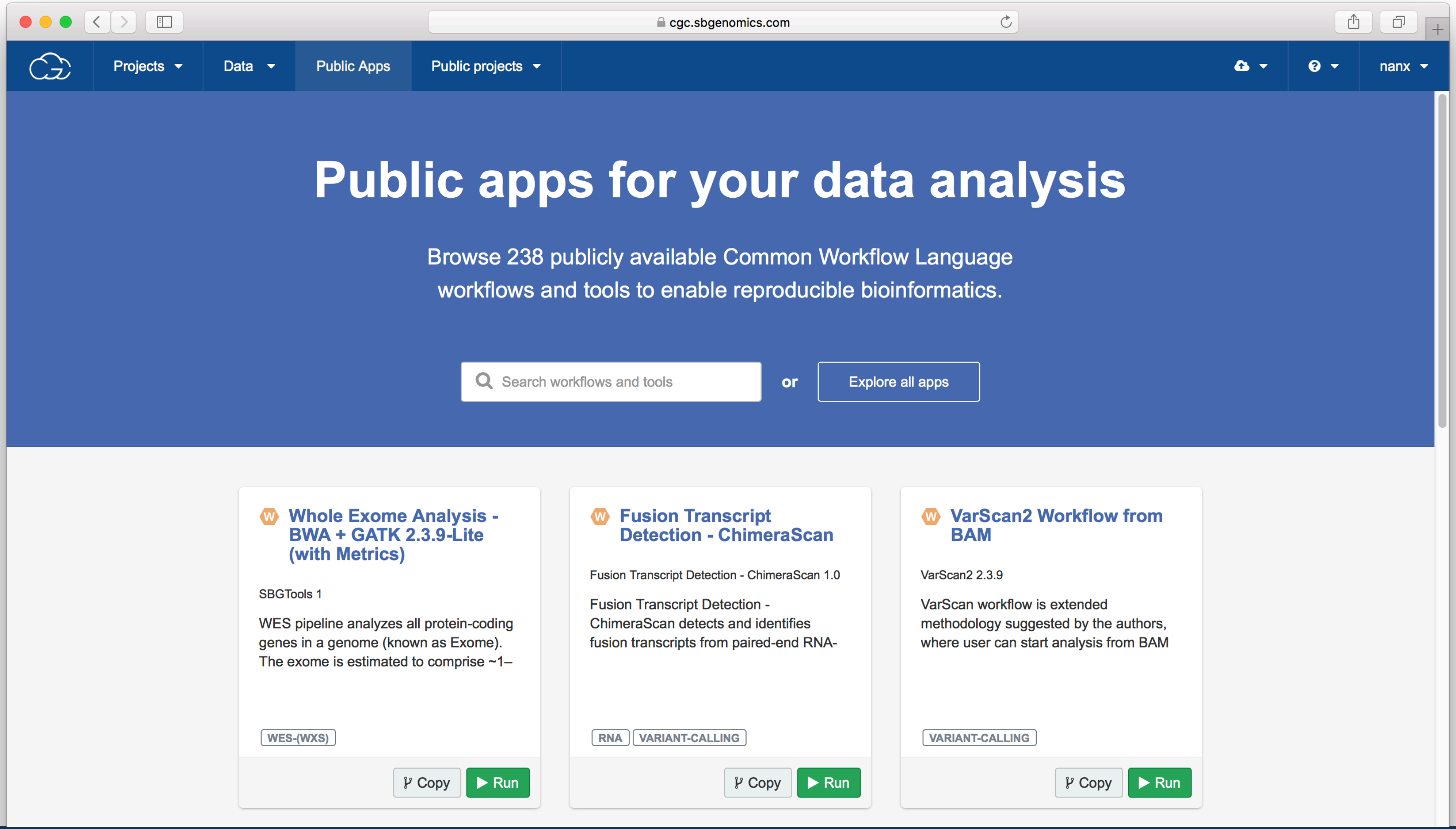
1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

IS THERE A REPRODUCIBILITY CRISIS?
7% Don't know
52% Yes, a significant crisis
3% No, there is no crisis
1,576 researchers surveyed

CANCER GENOMICS CLOUD (CGC)

- www.cancergenomicscloud.org
- Hundreds of automated analysis workflows for petabyte-scale data from The Cancer Genome Atlas



Public apps for your data analysis

Browse 238 publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

Whole Exome Analysis - BWA + GATK 2.3.9-Lite (with Metrics)

SBGTools 1

WES pipeline analyzes all protein-coding genes in a genome (known as Exome). The exome is estimated to comprise ~1–

Fusion Transcript Detection - ChimeraScan

Fusion Transcript Detection - ChimeraScan 1.0

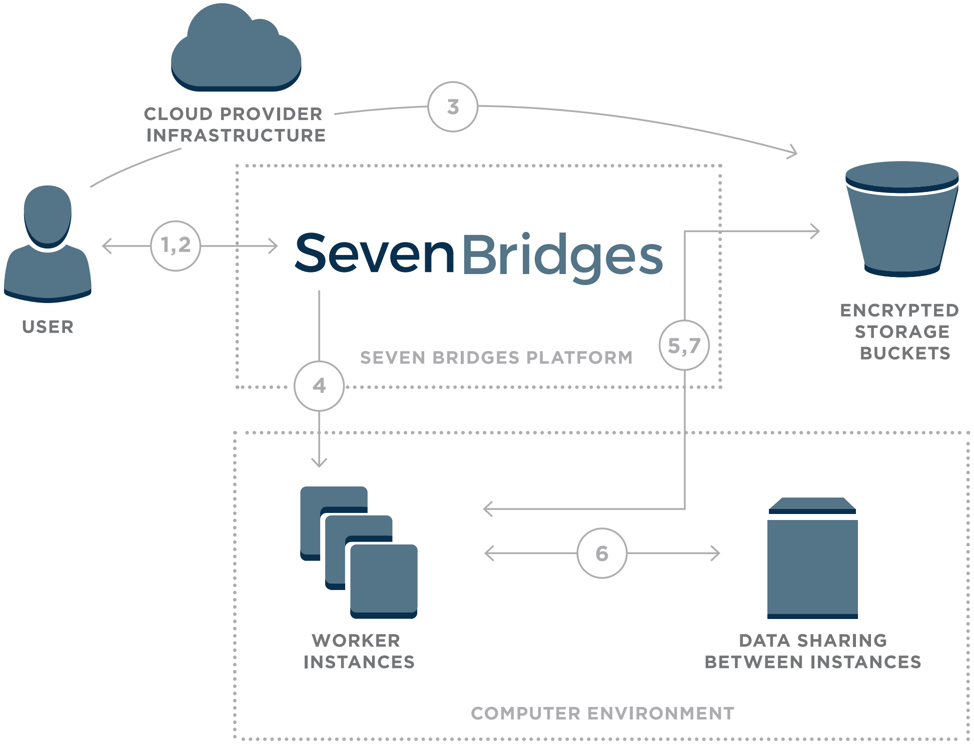
Fusion Transcript Detection - ChimeraScan detects and identifies fusion transcripts from paired-end RNA-

VarScan2 Workflow from BAM

VarScan2 2.3.9

VarScan workflow is extended methodology suggested by the authors, where user can start analysis from BAM

PRODUCT & ENGINEERING INNOVATIONS IN CGC


CLOUD PROVIDER INFRASTRUCTURE
3
USER
1,2
SevenBridges
SEVEN BRIDGES PLATFORM
4
5,7
ENCRYPTED STORAGE BUCKETS
WORKER INSTANCES
6
DATA SHARING BETWEEN INSTANCES
COMPUTER ENVIRONMENT

COMMON
WORKFLOW
LANGUAGE

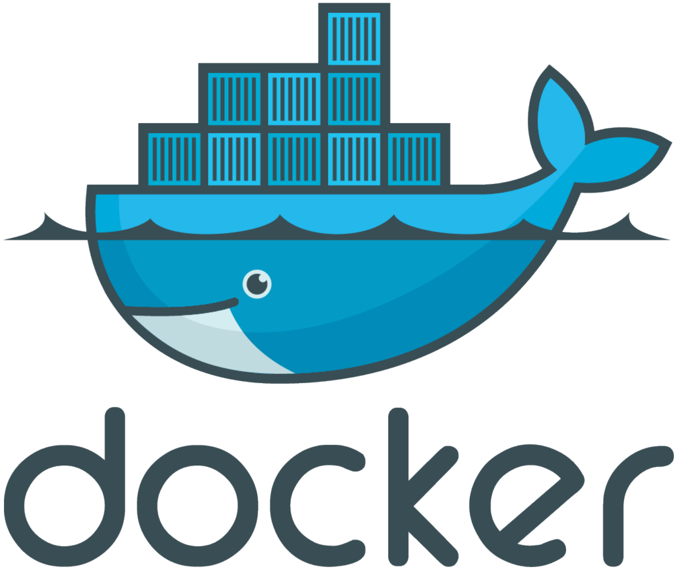
Rabix

CHALLENGE: OS-LEVEL REPRODUCIBILITY

How to ensure your reports are reproducible across time and environments, when the data, analysis tools, operating systems are all evolving?

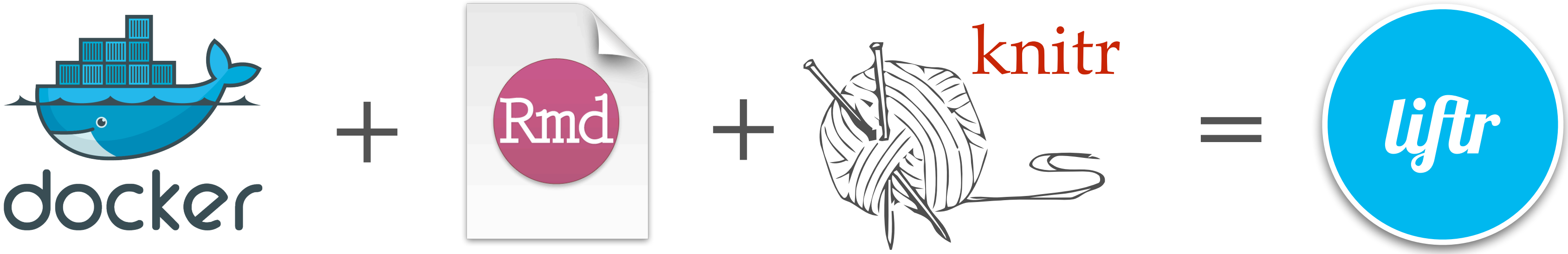
DOCKER

- Docker allows applications and their dependencies to be packaged into discrete runtime environments, called containers. Applications packaged in this way can be run from many diverse infrastructures.

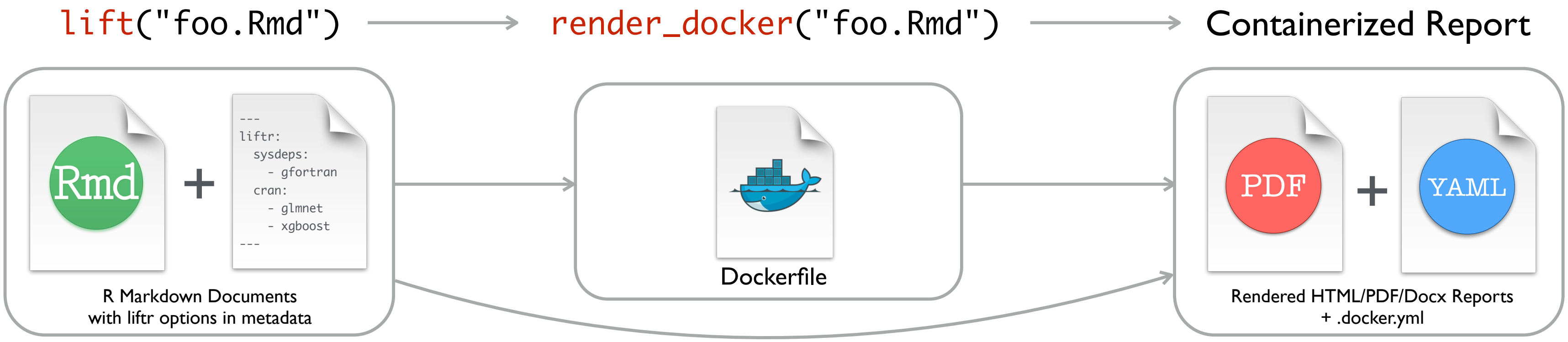


liftr

OS-level reproducibility & persistency for reports.



DOCKERIZE DOCUMENTS AS EASY AS 1-2-3



lift("foo.Rmd") → render_docker("foo.Rmd") → Containerized Report

```
---
liftr:
  sysdeps:
    - gfortran
  cran:
    - glmnet
    - xgboost
---
```

R Markdown Documents
with liftr options in metadata

Dockerfile

Rendered HTML/PDF/Docx Reports
+ .docker.yml

liftr extends the R Markdown metadata format, introducing additional options for containerizing and rendering reports.

By running lift() on the RMD file, liftr parses the metadata fields appeared in the R Markdown document; then generates the Dockerfile.

By running render_docker(), liftr will build the Docker image, run the container, and render the R Markdown document.

DOCKERIZE DOCUMENTS AS EASY AS 1-2-3

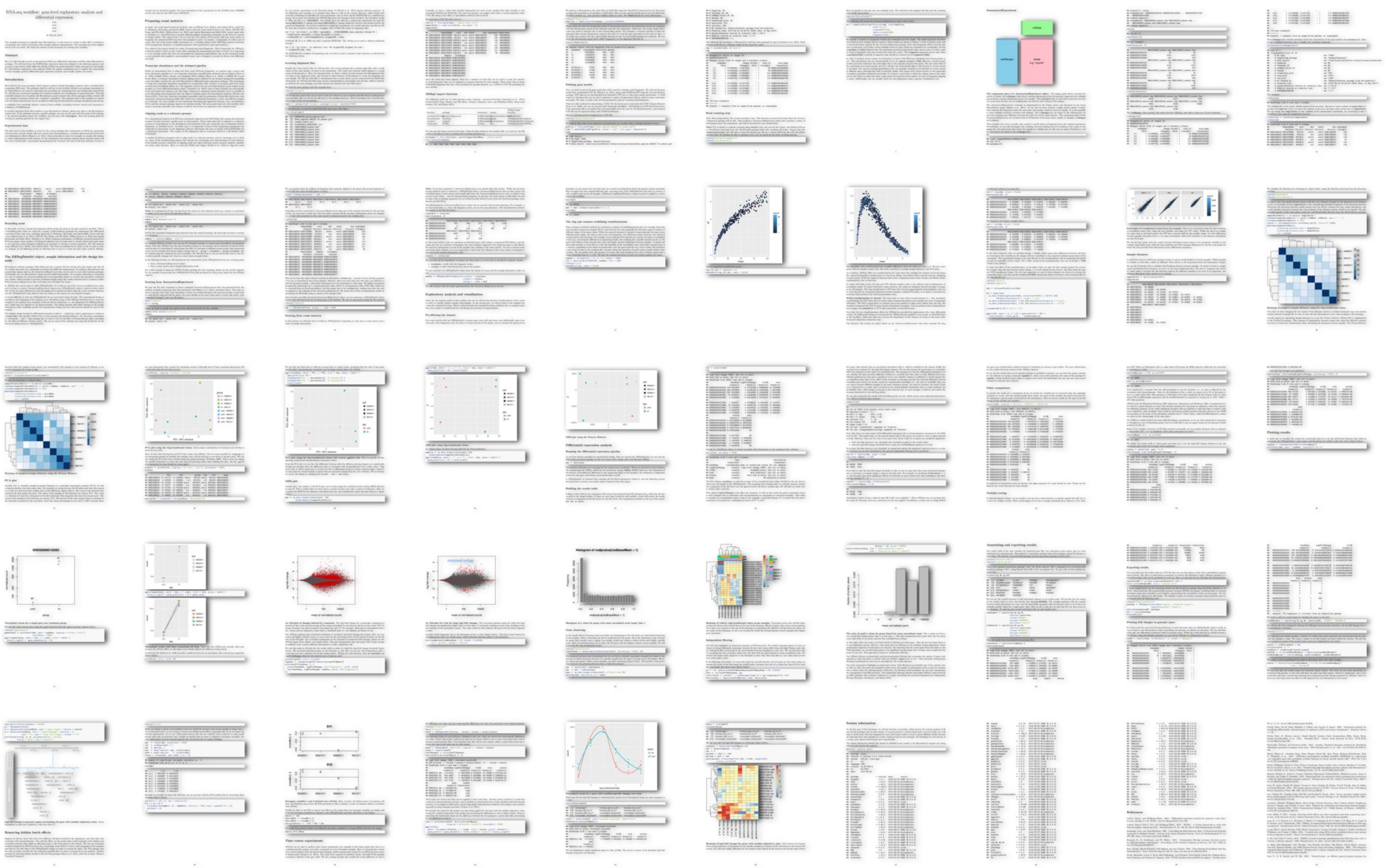
```
library("liftr")
input = "demo.Rmd"

lift(input)           # Generate Dockerfile
render_docker(input)  # Render report with Docker

purge_image(input)    # Clean up Docker image
push_image(input)     # Push image to registry (devel)
```

DEMO: RNA-SEQ DATA ANALYSIS

Example workflow from Bioconductor.org

- RNA-Seq: biotechnology for measuring the expression of genes. It can help identify potential key genes in cancer.
- TBs of RNA-Seq data are generated. Computational tools and workflows are developed to analyze such data.
- We need to ensure such reports are reproducible through time, when datasets, analysis tools are both evolving.
- Code available from: bit.ly/liftrdemo

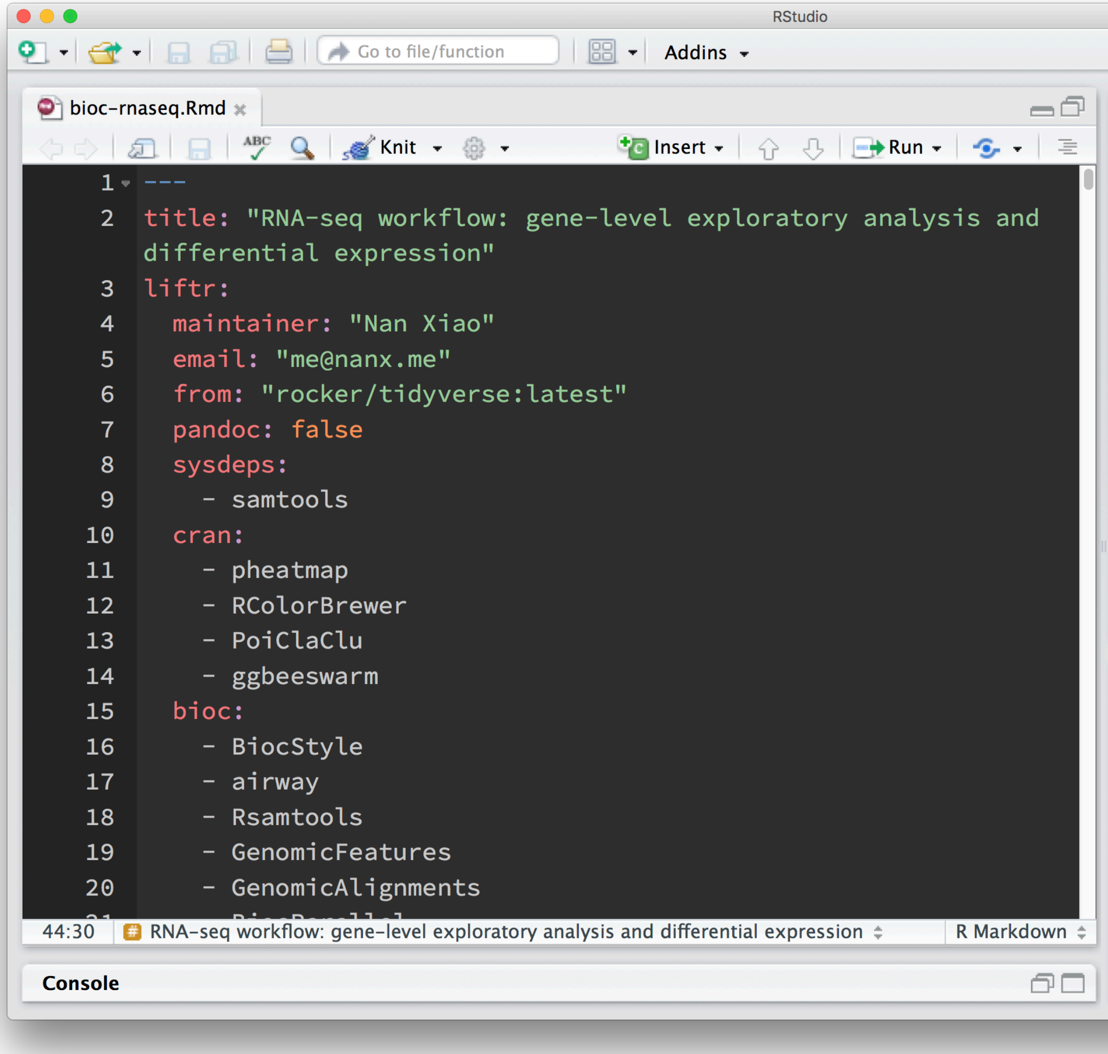
COMPLEXITY IN DEPENDENCY



STEP 1

Add liftr metadata to the R Markdown document:

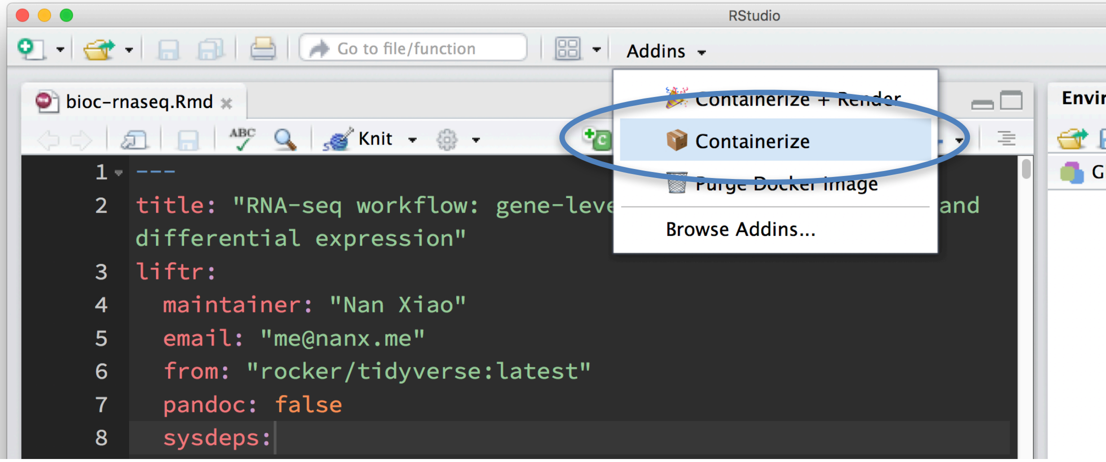
- Base image
- System dependencies
- Package dependencies
- …



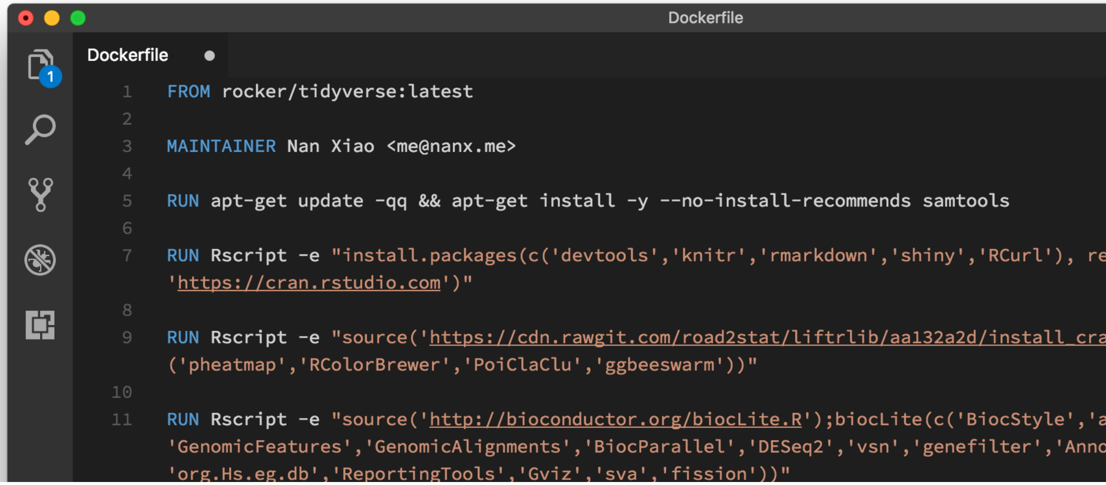
```
---
title: "RNA-seq workflow: gene-level exploratory analysis and
differential expression"
liftr:
  maintainer: "Nan Xiao"
  email: "me@nanx.me"
  from: "rocker/tidyverse:latest"
  pandoc: false
  sysdeps:
    - samtools
  cran:
    - pheatmap
    - RColorBrewer
    - PoiClaClu
    - ggbeeswarm
  bioc:
    - BiocStyle
    - airway
    - Rsamtools
    - GenomicFeatures
    - GenomicAlignments
    - BiocParallel
```

STEP 2

Generate Dockerfile with liftr::lift



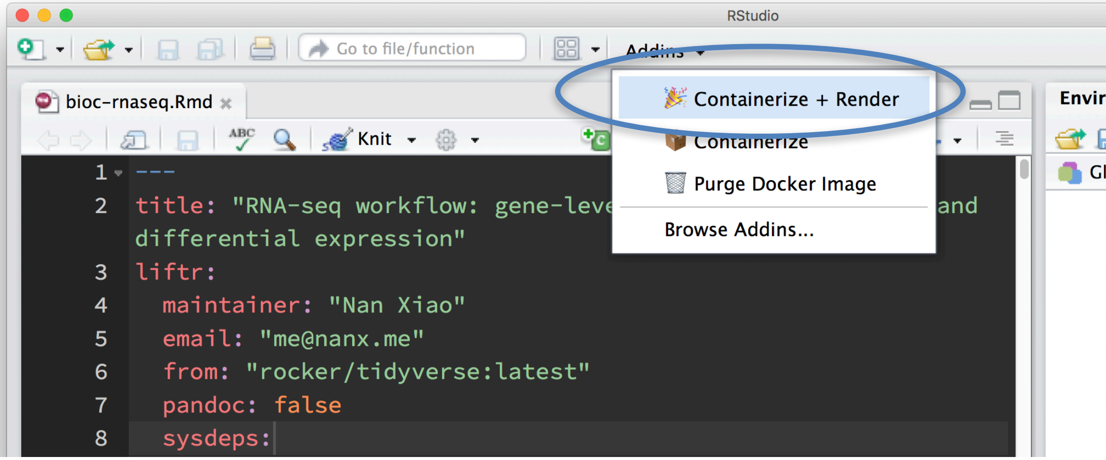
```
---
title: "RNA-seq workflow: gene-leve                and
differential expression"
liftr:
  maintainer: "Nan Xiao"
  email: "me@nanx.me"
  from: "rocker/tidyverse:latest"
  pandoc: false
  sysdeps:
```



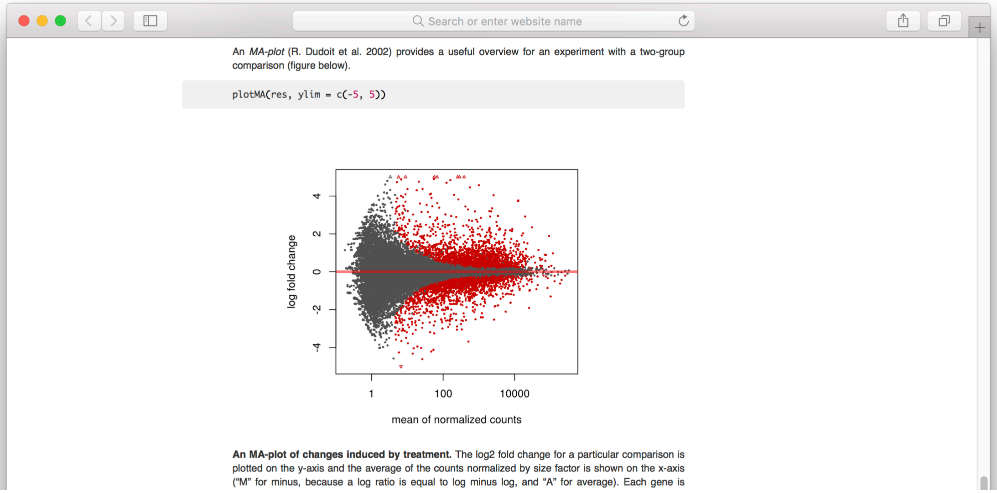
```
FROM rocker/tidyverse:latest

MAINTAINER Nan Xiao <me@nanx.me>

RUN apt-get update -qq && apt-get install -y --no-install-recommends samtools

RUN Rscript -e "install.packages(c('devtools','knitr','rmarkdown','shiny','RCurl'), re
'https://cran.rstudio.com')"

RUN Rscript -e "source('https://cdn.rawgit.com/road2stat/liftrlib/aa132a2d/install_cra
('pheatmap','RColorBrewer','PoiClaClu','ggbeeswarm'))"

RUN Rscript -e "source('http://bioconductor.org/biocLite.R');biocLite(c('BiocStyle','a
'GenomicFeatures','GenomicAlignments','BiocParallel','DESeq2','vsn','genefilter','Anno
'org.Hs.eg.db','ReportingTools','Gviz','sva','fission'))"
```

STEP 3

- liftr::render_docker will build the Docker image, run the container, and render into PDF/HTML/DOCX.
- Re-compile: cached Docker image layers are used to improve speed.
- Remove the used image, or push to Docker registry.



```
---
title: "RNA-seq workflow: gene-leve and
differential expression"
liftr:
  maintainer: "Nan Xiao"
  email: "me@nanx.me"
  from: "rocker/tidyverse:latest"
  pandoc: false
  sysdeps:
```



An *MA-plot* (R. Dudoit et al. 2002) provides a useful overview for an experiment with a two-group comparison (figure below).

```
plotMA(res, ylim = c(-5, 5))
```

An MA-plot of changes induced by treatment. The log2 fold change for a particular comparison is plotted on the y-axis and the average of the counts normalized by size factor is shown on the x-axis ("M" for minus, because a log ratio is equal to log minus log, and "A" for average). Each gene is

FUTURE WORKS

We aim to expand the R Markdown tool chain by exploring the next frontier: system-level reproducibility, and democratize reproducible report creation/sharing.

To achieve this, we need:

- Standard renderers + independent YAML configuration file
- Better IDE support (RStudio Addins)
- Better on-boarding experience: automatic dependency parsing
- Cloud-based rendering and containerization services for dynamic documents

Q & A

Visit liftr.me for more info

Contact: me@nanx.me