



Persistent Reproducible Reporting

# Nan Xiao

Genomic Data Scientist,  
Seven Bridges



# Reproducible Research



# R Markdown + knitr to the rescue

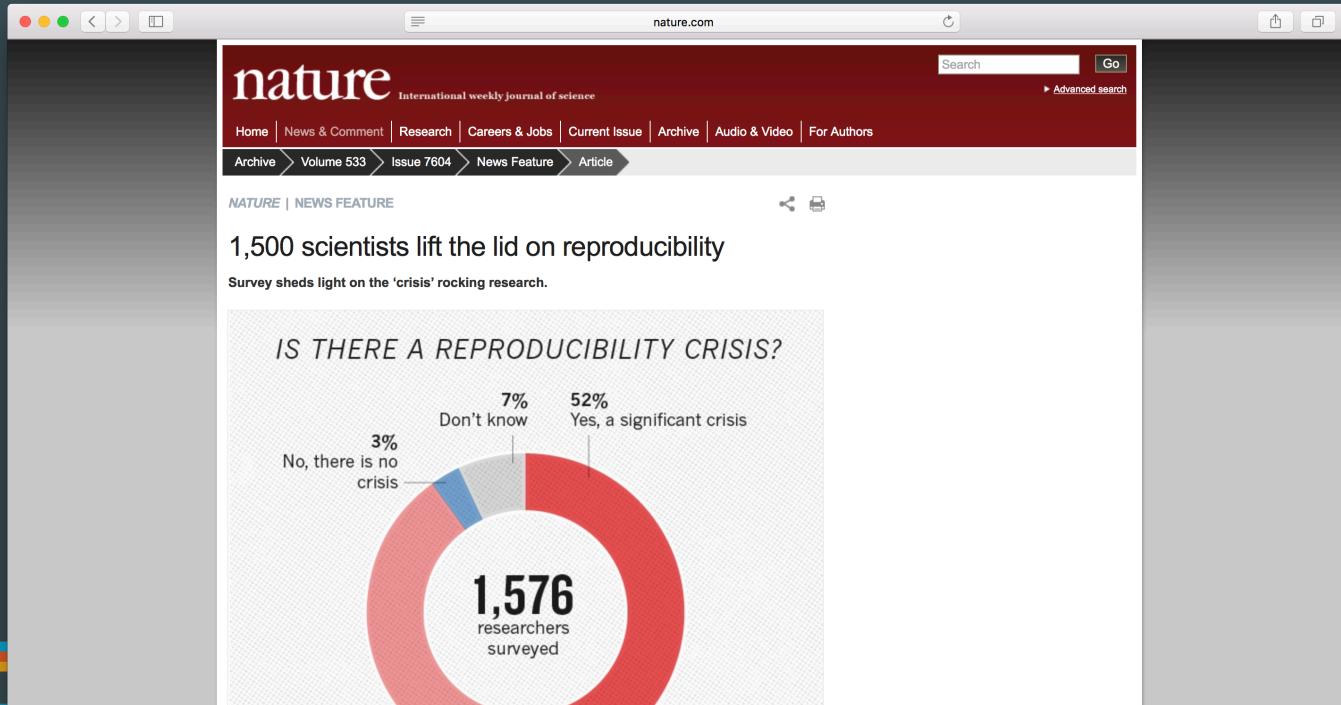


+



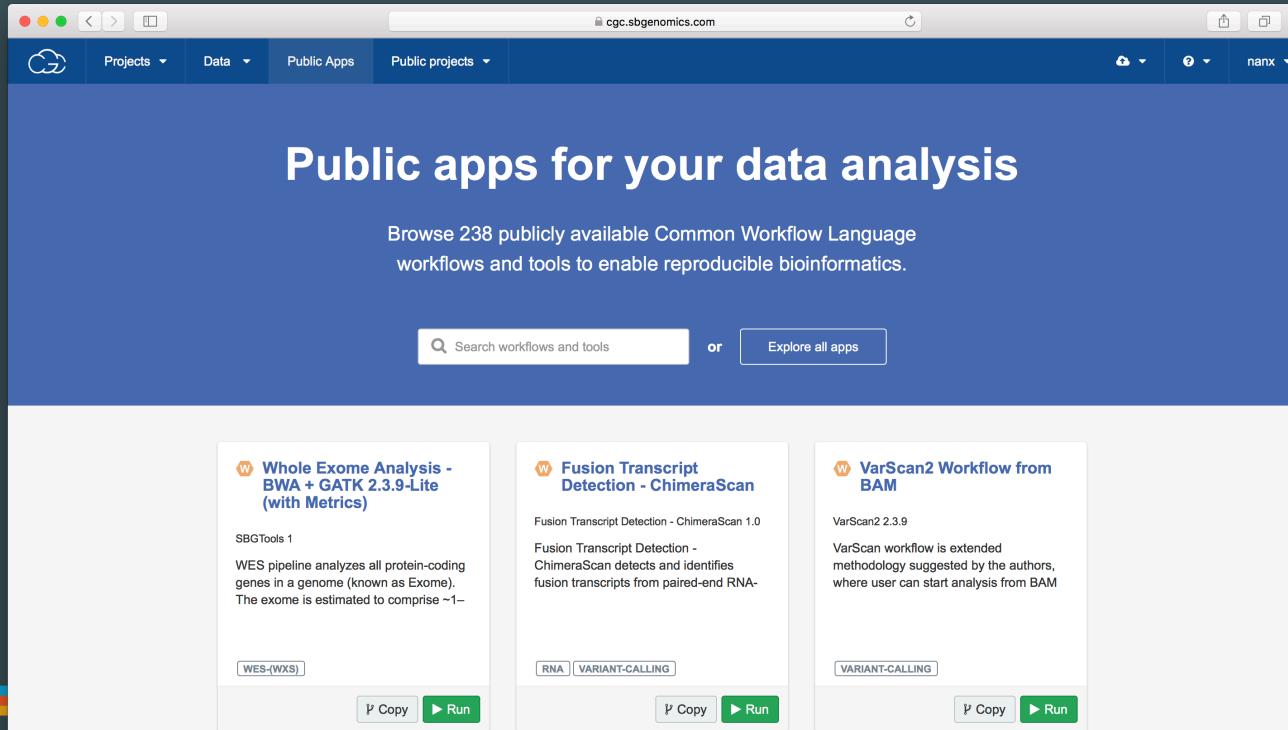
# Reproducibility

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



# Cancer Genomics Cloud

[www.cancergenomicscloud.org](http://www.cancergenomicscloud.org)



The screenshot shows the 'Public Apps' section of the CGC website. At the top, there's a search bar with 'Search workflows and tools' and an 'Explore all apps' button. Below the search bar, three workflow cards are displayed:

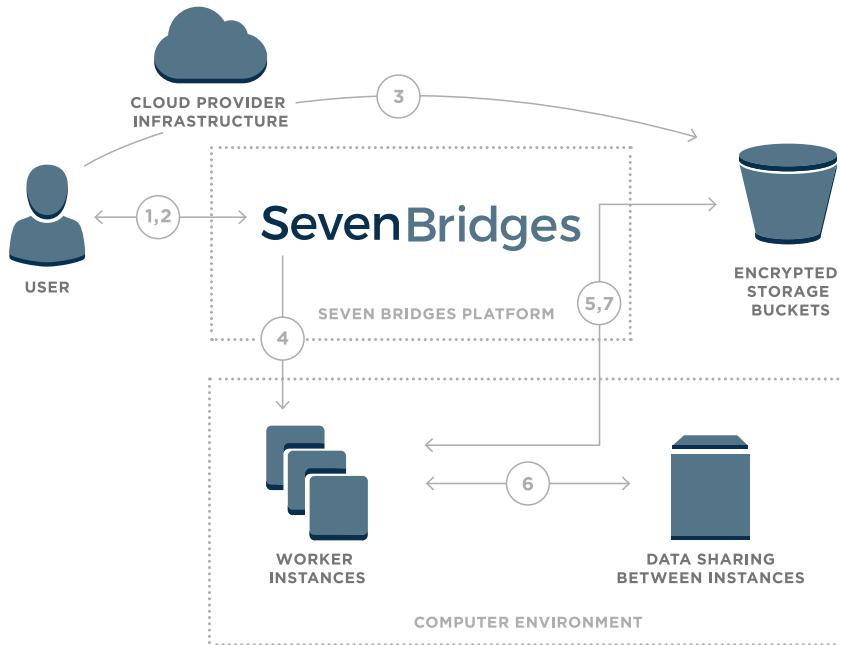
- Whole Exome Analysis - BWA + GATK 2.3.9-Lite (with Metrics)**  
SBGTools 1  
WES pipeline analyzes all protein-coding genes in a genome (known as Exome). The exome is estimated to comprise ~1%
- Fusion Transcript Detection - ChimeraScan**  
Fusion Transcript Detection - ChimeraScan 1.0  
Fusion Transcript 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

Each card has a 'WES-(WXS)' button at the bottom left, and 'Copy' and 'Run' buttons at the bottom right.

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



# Product & Tech Innovations in CGC

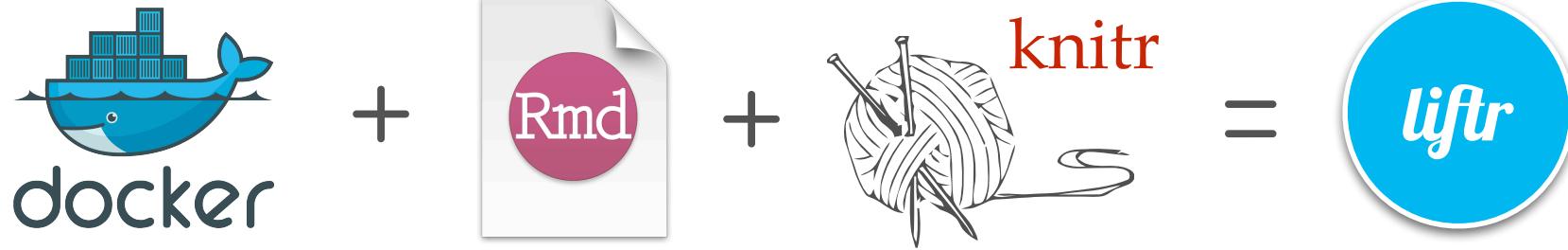


Rabix

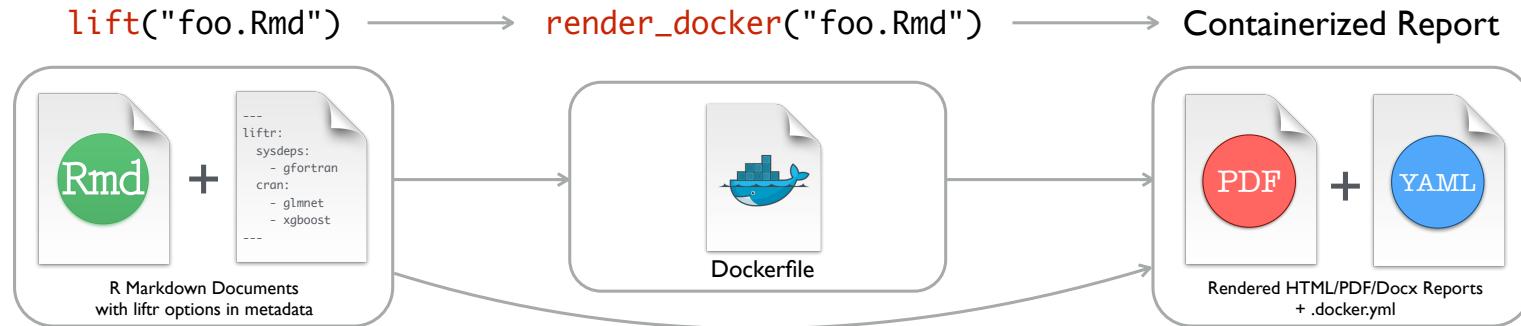


# liftr

OS-level reproducibility & persistency for reports.



# Dockerize documents as easy as 1-2-3



`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

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

# Step 1

Add liftr metadata to the R Markdown document:  
base image, system dependencies, package dependencies, etc.



RStudio

bioc-rnaseq.Rmd

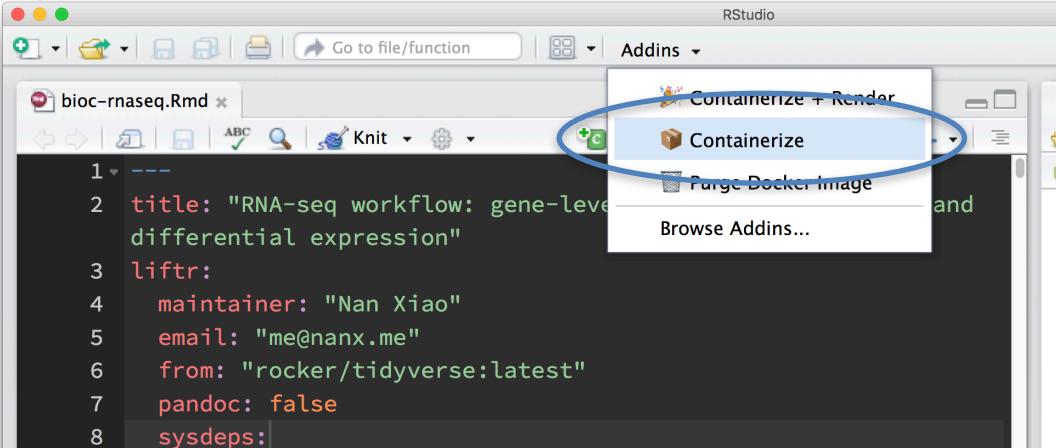
```
1 ---  
2 title: "RNA-seq workflow: gene-level exploratory analysis and  
3 differential expression"  
4 liftr:  
5   maintainer: "Nan Xiao"  
6   email: "me@nanx.me"  
7   from: "rocker/tidyverse:latest"  
8   pandoc: false  
9   sysdeps:  
10    - samtools  
11   cran:  
12    - pheatmap  
13    - RColorBrewer  
14    - PoiClaClu  
15    - ggbeeswarm  
16   bioc:  
17    - BiocStyle  
18    - airway  
19    - Rsamtools  
20    - GenomicFeatures  
21    - GenomicAlignments
```

44:30 RNA-seq workflow: gene-level exploratory analysis and differential expression R Markdown

Console

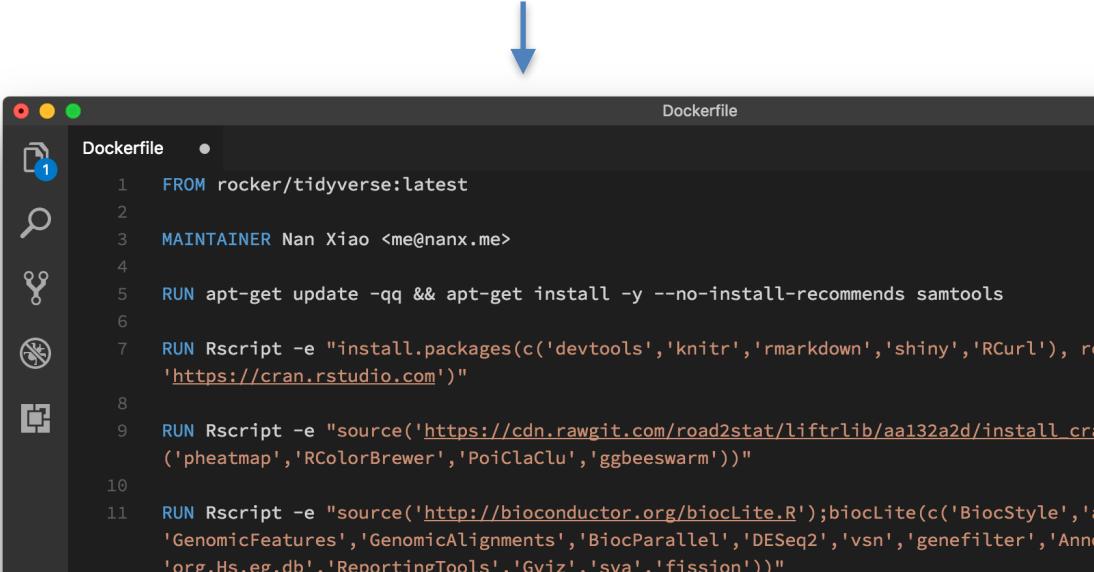
# Step 2

Use `liftr::lift` to generate Dockerfile



The screenshot shows the RStudio interface with a file named "bioc-rnaseq.Rmd" open. A context menu is displayed from the "Addins" dropdown, with the "Containerize" option highlighted. The menu also includes "Purge Docker Image" and "Browse Addins...".

```
1 ---  
2 title: "RNA-seq workflow: gene-level differential expression"  
3 liftr:  
4   maintainer: "Nan Xiao"  
5   email: "me@nanx.me"  
6   from: "rocker/tidyverse:latest"  
7   pandoc: false  
8   sysdeps:
```



An arrow points from the "Containerize" menu in the RStudio interface down to a screenshot of a Dockerfile. The Dockerfile contains several commands to set up a container environment, including installing packages and running R scripts to install specific bioconductor packages.

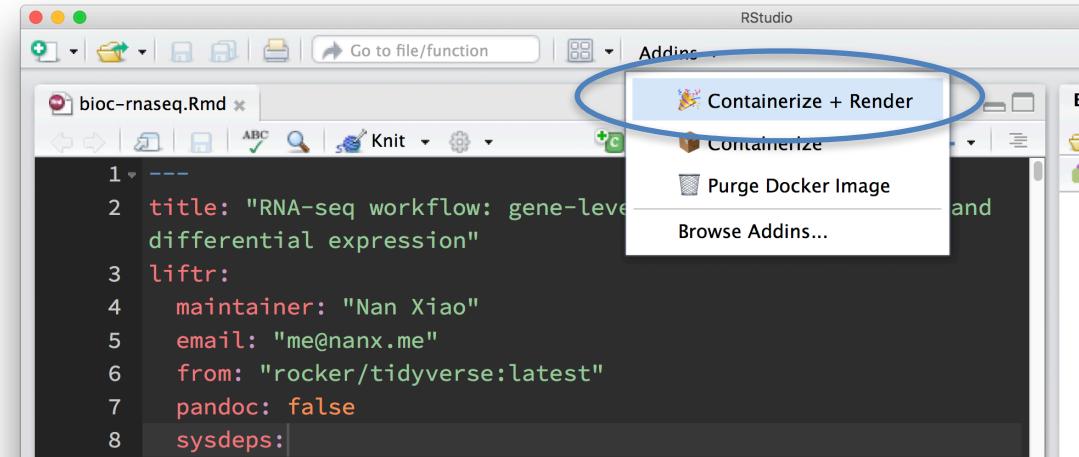
```
FROM rocker/tidyverse:latest  
  
MAINTAINER Nan Xiao <me@nanx.me>  
  
RUN apt-get update -qq && apt-get install -y --no-install-recommends samtools  
  
RUN Rscript -e "install.packages(c('devtools','knitr','rmarkdown','shiny','RCurl'),  
  'https://cran.rstudio.com')"  
  
RUN Rscript -e "source('https://cdn.rawgit.com/road2stat/liftrlib/aa132a2d/install_cra  
('pheatmap','RColorBrewer','PoiClaClu','ggbeeswarm'))"  
  
RUN Rscript -e "source('http://bioconductor.org/biocLite.R');biocLite(c('BiocStyle',  
  'GenomicFeatures','GenomicAlignments','BiocParallel','DESeq2','vsn','genefilter','Anno  
 'org.Hs.eg.db','ReportingTools','Gviz','sva','fission'))"
```

# Step 3

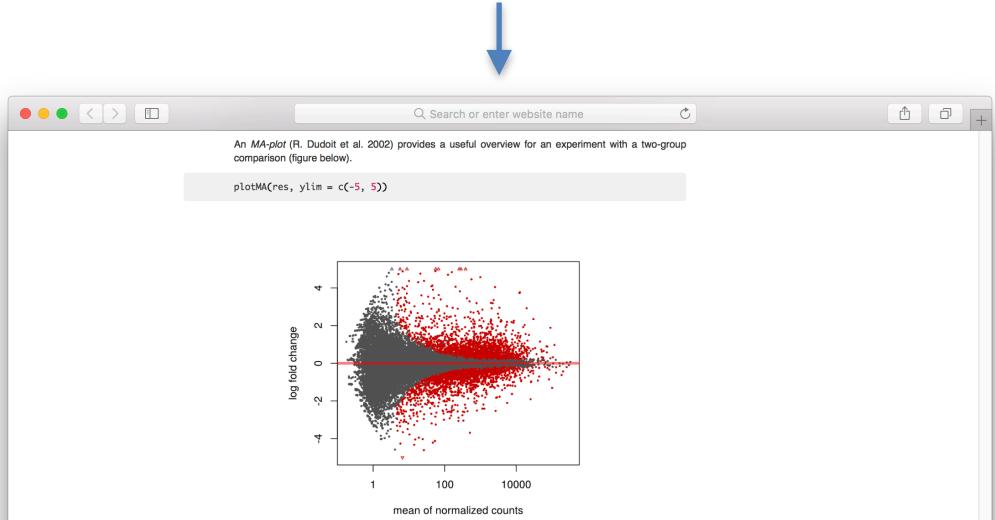
`liftr::render_docker`  
will build the image, run the  
container, and render into  
PDF/HTML/Docx.

Re-compilation: cached  
image layers are used to  
improve speed.

Remove the used image, or  
push to registry.



```
1 ---  
2 title: "RNA-seq workflow: gene-level  
differential expression"  
3 liftr:  
4   maintainer: "Nan Xiao"  
5   email: "me@nanx.me"  
6   from: "rocker/tidyverse:latest"  
7   pandoc: false  
8   sysdeps:
```



# Future works

- Cloud-based rendering and containerization services for dynamic documents
- Democratize reproducible report creation/sharing

# Thank You!

[liftr.me](http://liftr.me)

@road2stat

#dockercon #liftr

