SevenBridges

Deploying your own tools on the CGC

with Docker and CWL

Gaurav Kaushik, PhD
Scientific Program Manager

As data has grown, so has the number of tools to analyze it

11,000+ -omics data analysis tools*
(each with many versions)

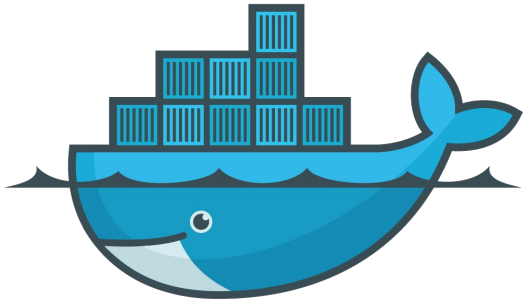
50+ used in a single TCGA marker paper

*omictools.com

Can we make tools portable so they're easily deployed on the cloud?

(sì possiamo!)

Scalable, Reproducible, Portable Bioinformatics with

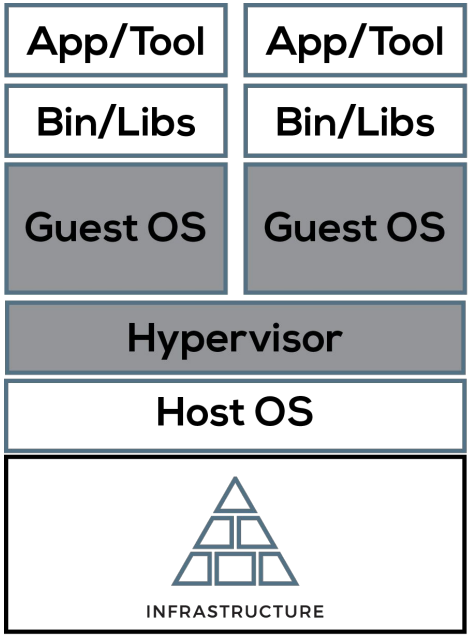


+

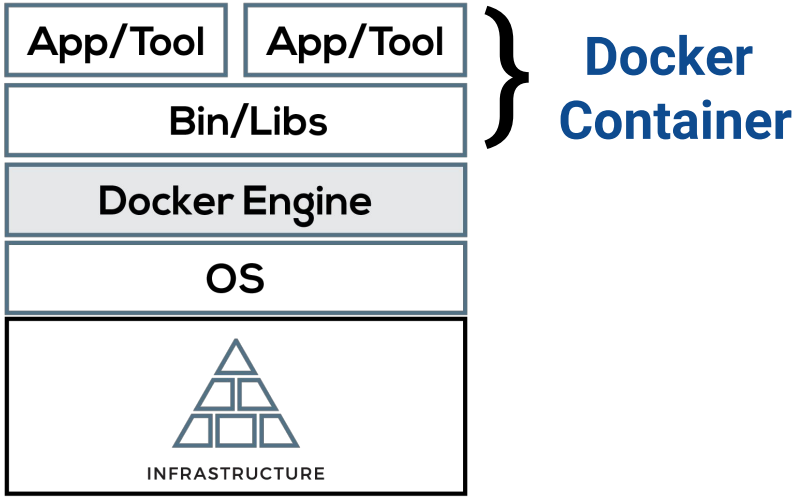
COMMON
WORKFLOW
LANGUAGE

Docker

lightweight and portable



VM



DOCKER

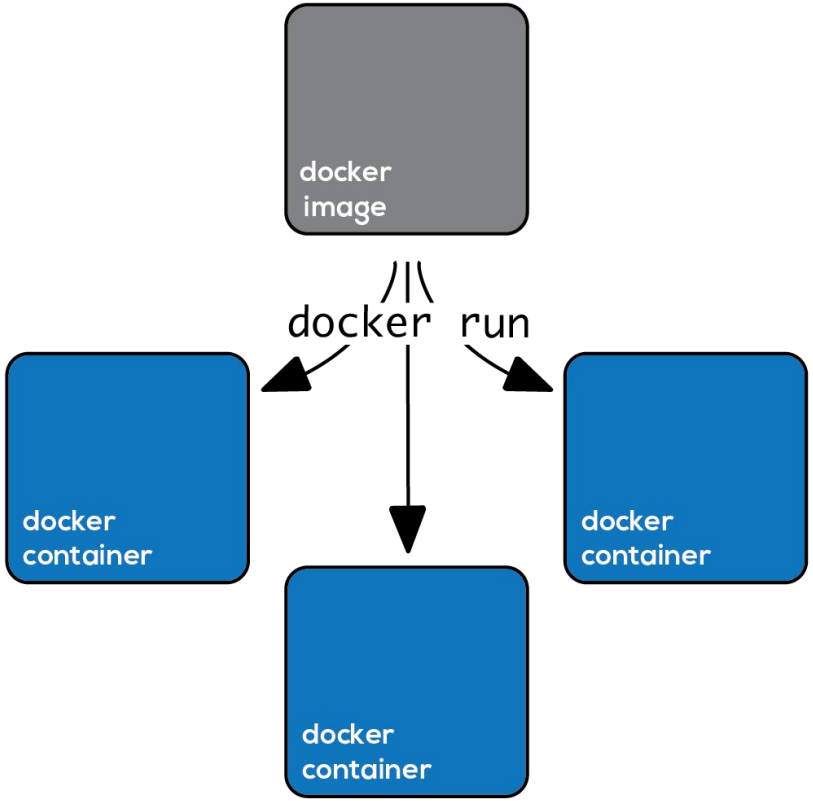
Images + Containers

Images are read-only
Containers are mutable instances of an image

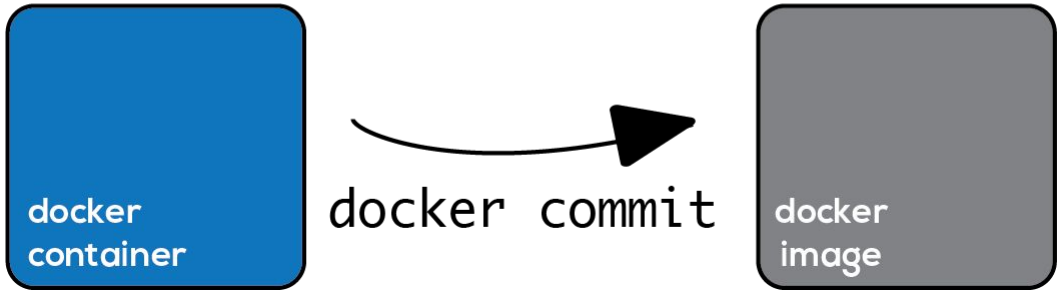
read-only

read+write

Multiple containers can be instantiated from an image



To persist changes, commit a container to a new image



```
gaurav~:$ docker images
REPOSITORY                          TAG       IMAGE ID        CREATED        VIRTUAL SIZE
gauravkaushik/freebayes             latest    sha256:653cc    5 hours ago    941.4 MB
ubuntu                              latest    sha256:e17b5    5 days ago     188 MB
rfranklin/pythondev                 latest    sha256:f0ce0    8 days ago     1.868 GB
rfranklin/rstatsdev                 latest    sha256:a6ffe    8 days ago     2.962 GB
images.sbgenomics.com/gaurav/jellyfish  2.2.4  sha256:67522   3 weeks ago    422.6 MB
gaurav~:$
```

To open an ubuntu environment:

$ docker run ubuntu

...and that's it!

Using Docker

There are three ways to get a Docker container with your software:

1. Use a public container (e.g. Bioconductor, Anaconda)
2. Build on top of a public container
3. Build from “scratch”

Using Docker

If you build your own you can do it

Interactively or using a *Dockerfile*

Using Docker Interactively

docker pull repo/container

docker run repo/container

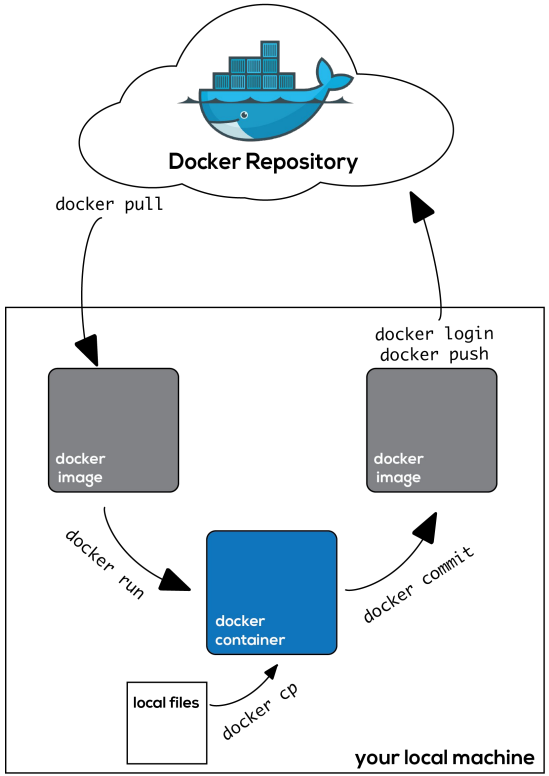
Using Dockerfiles

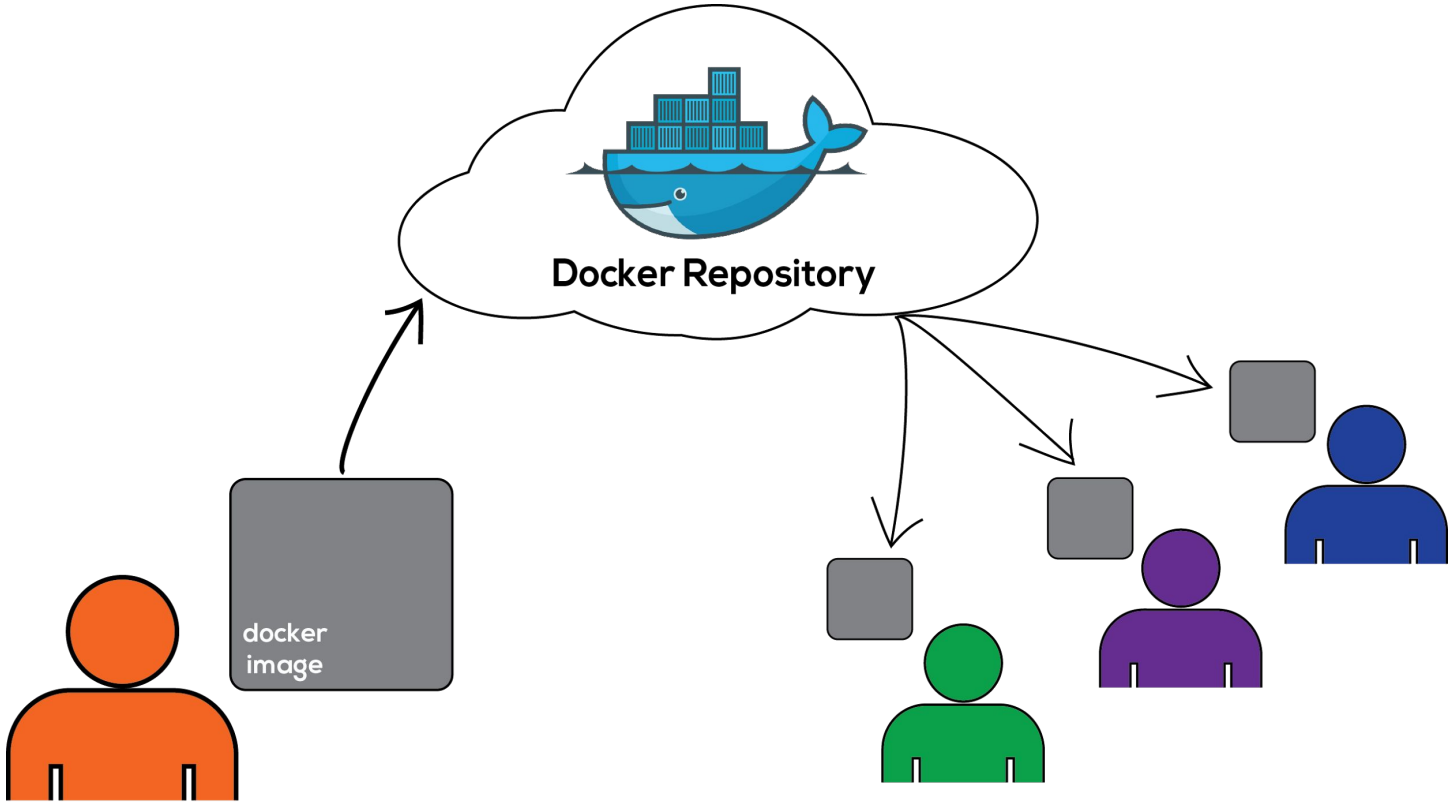
docker build -t repo/container .

```
# How to use Dockerfiles

# Start with a base image
FROM ubuntu

MAINTAINER "Gaurav Kaushik"

# Update the container
RUN apt-get update --yes
RUN apt-get install vim gcc --yes

# Set CMD to fire up the terminal
CMD ["/bin/bash"]
```

Make it usable on the CGC

docker push repo/container





Now that I “Dockerized” my software

how do I run it?

COMMON
WORKFLOW
LANGUAGE

a standard for simply-described, reproducible analyses

A team effort since 2014

Global Alliance for Genomic Health (GA4GH)

Institute for Systems Biology

Galaxy

Curoverse

Wellcome Trust Sanger Institute

Institut Pasteur

University of California Santa Cruz

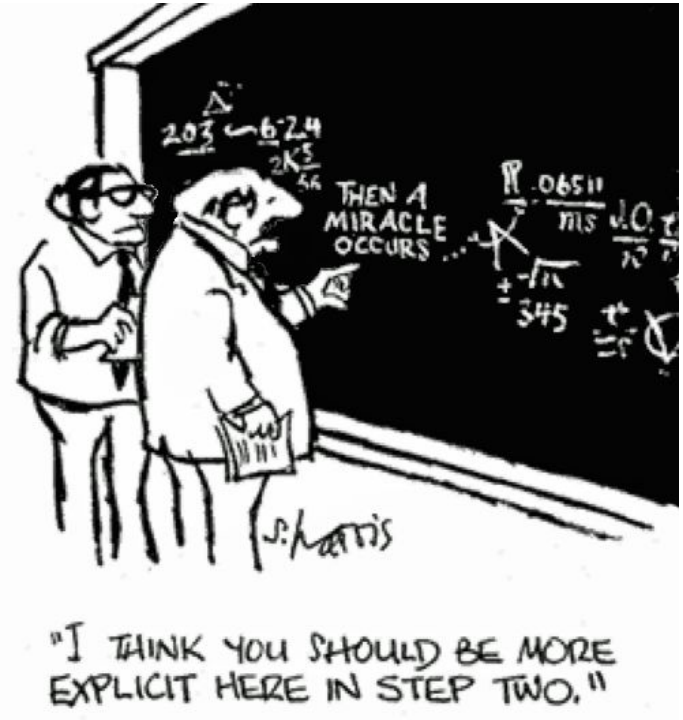
Intel

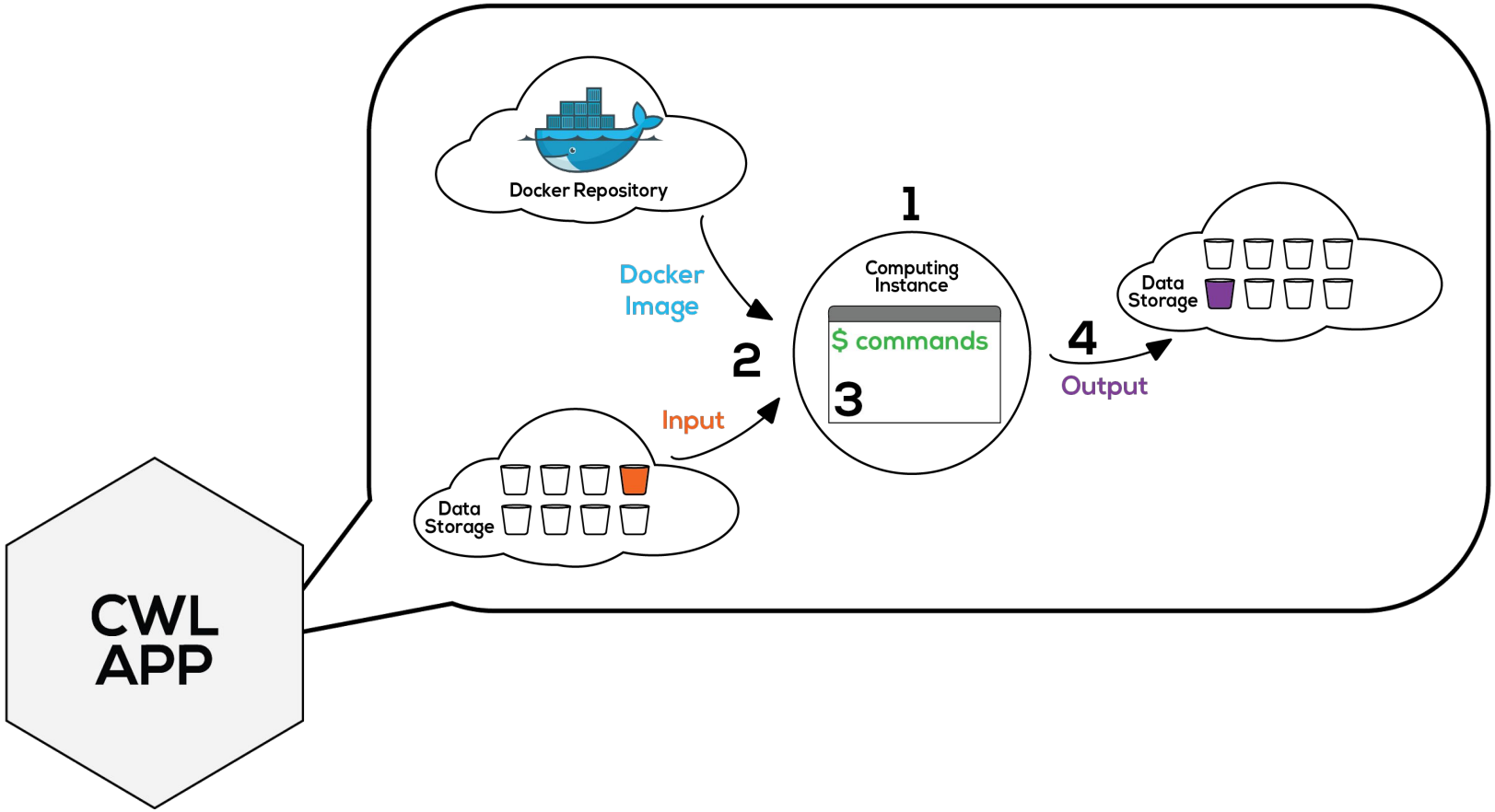
...and 20+ more institutions and commercial partners

Reproducibility shouldn't be hard

To reproduce most findings right now:

- Email the authors/old lab members
- Build each original tool from source
- Replicate the entire pipeline
- Move the data around and/or reformat





How do I learn CWL?

The syntax and excellent resources can be found at commonwl.org

BUT with the Seven Bridges Software Development Kit,
you can create tools and chain them into workflows interactively.

The Seven Bridges SDK will create the CWL code for you
so you can get up and running on the platform **more quickly and easily**.

Rabix from Seven Bridges

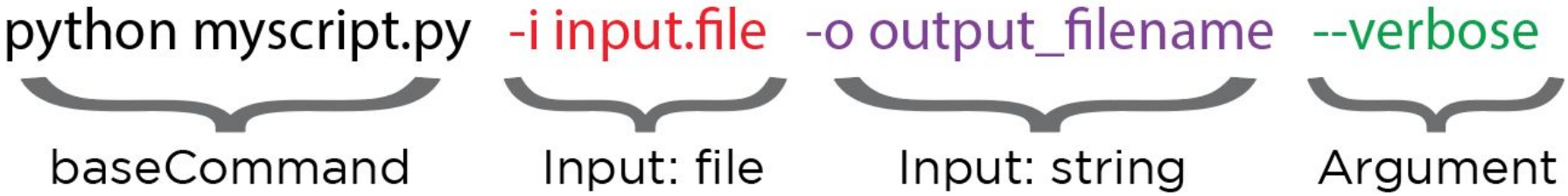
Software Development Kit for **composing** and **executing**

Common Workflow Language-described applications.

Each application has:

1. a Docker container with the software+dependencies to run it
2. a CWL description of how to run software in the container

Wrapping a tool



Wrapping a tool

Case #1: *grep*

***G**lobal **R**egular **E**xpression **P**rint*

Case #1: *grep*

grep

baseCommand: grep

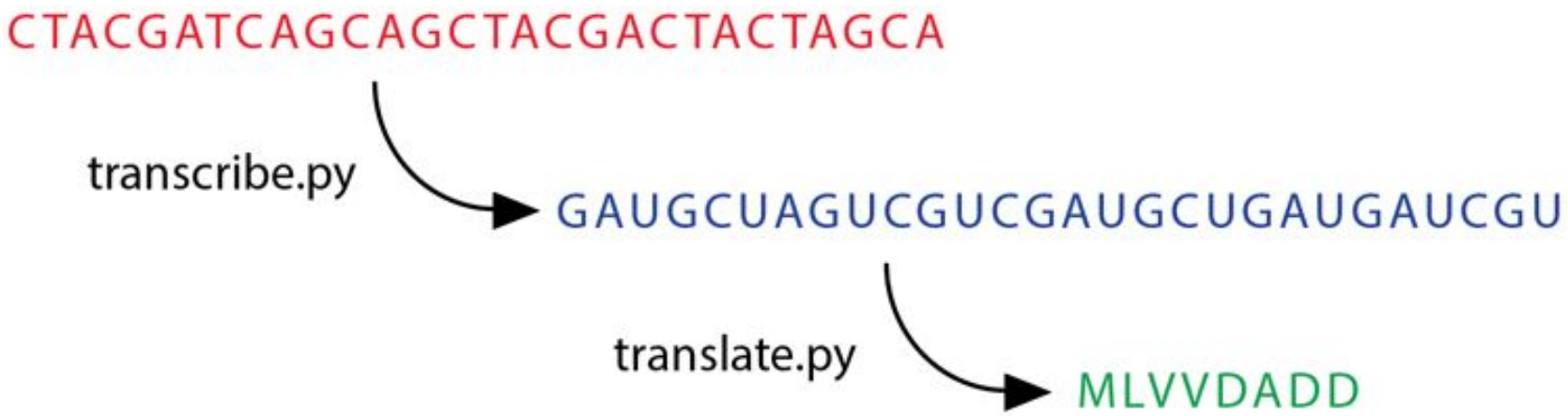
Inputs:

- string: Pattern
- File

Outputs: stdout

```
gaurav~:$ cat test.txt
hello
this file
is for
testing1
testin1g
testi1ng
test1ing
tes1ting
te1sting
t1esting
1testing
gaurav~:$ grep "test" test.txt
testing1
testin1g
testi1ng
test1ing
1testing
gaurav~:$
```

Case #2: dna2protein



Case #2: dna2protein

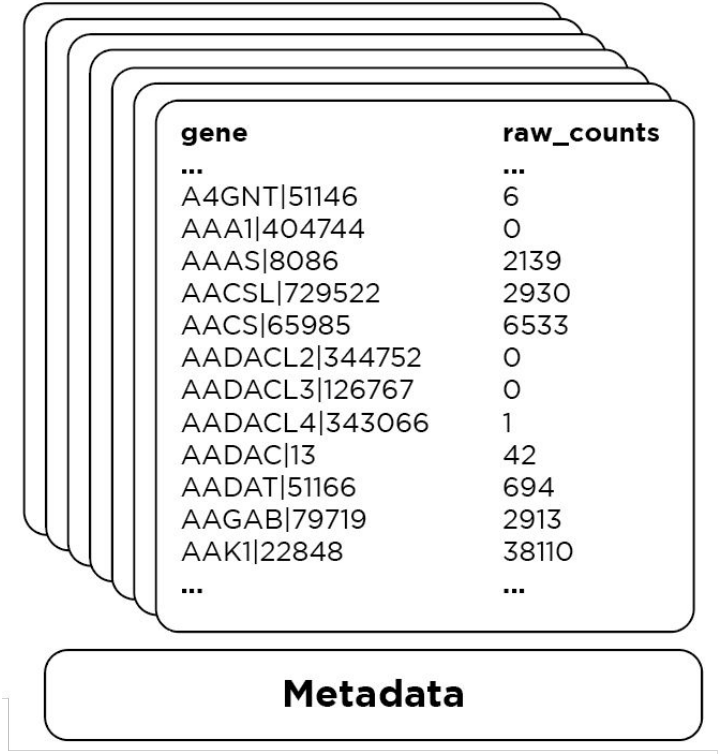
Transcribe
baseCommand: python transcribe.py
Inputs: -d dna.txt
Outputs: stdout

Translate
baseCommand: python translate.py
Inputs: -r rna.txt
Outputs: stdout

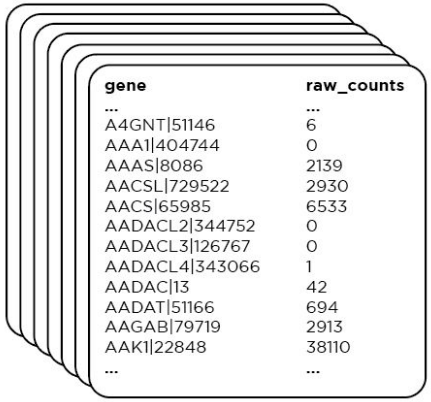
Case #3:

Differential Expression

Case #3: Differential Expression



Case #3: Differential Expression



gene	raw_counts
...	...
A4GNT|51146	6
AAA1|404744	0
AAAS|8086	2139
AACSL|729522	2930
AACS|65985	6533
AADACL2|344752	0
AADACL3|126767	0
AADACL4|343066	1
AADAC|13	42
AADAT|51166	694
AAGAB|79719	2913
AAK1|22848	38110
...	...

gene	sample1	sample2	sample3	...
...	...	...	...	
A4GNT|51146	0	0	2	
AAA1|404744	1	0	0	
AAAS|8086	1004	831	1161	
AACSL|729522	1	0	2	
AACS|65985	1001	2329	1209	
AADACL2|344752	0	0	0	
AADACL3|126767	0	0	0	
AADACL4|343066	1	0	0	
AADAC|13	12	19	13	
AADAT|51166	452	52	300	
AAGAB|79719	1678	1272	1315	
AAK1|22848	593	908	590	
...	...	...	...	

Metadata

sample	gender	sample_type	...
sample1	Female	Primary Tumor	
sample2	Female	Solid Tissue Normal	
sample3	Female	Primary Tumor	
...	...	...	

Case #3: Differential Expression

Gene Expression Munger

baseCommand: python munger.py

Inputs:

    -r index_file

    -o output_filename

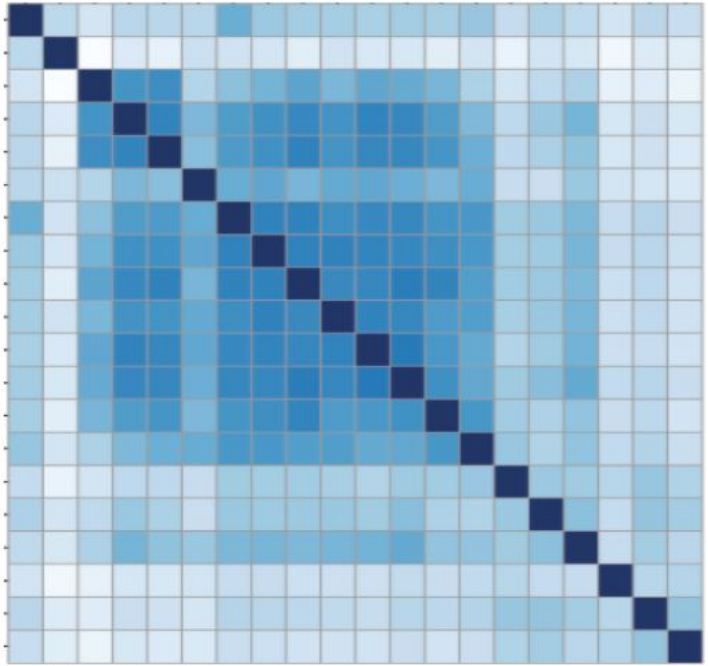
Outputs:

    gene.csv

    metadata.csv

Case #3: Differential Expression

gene	padj
MMP1|4312	1.0e-61
CST1|1469	1.4e-60
COL10A1|1300	4.3e-52
LALBA|3906	7.7e-51
CSN2|1447	2.2e-44
WIF1|11197	2.2e-44
MMP13|4322	2.7e-43
...	...



Case #3: Differential Expression

Differential Expression
baseCommand: Rscript diff.R
Inputs:
    gene.csv
    metadata.csv
Outputs:
    report.csv
    plots.pdf

For more information

Homepage: cancergenomicscloud.org

Knowledge Center: docs.cancergenomicscloud.org

Request additional funding