Transposable elements and Scientific cloud computing

Group Meeting - Jan 2018

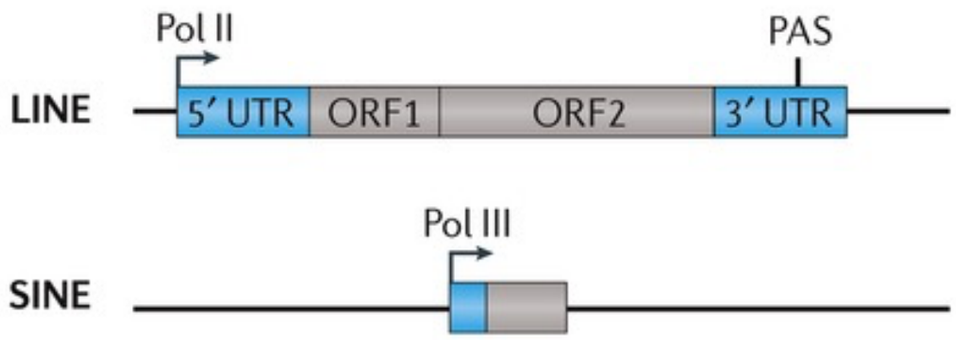
Agenda

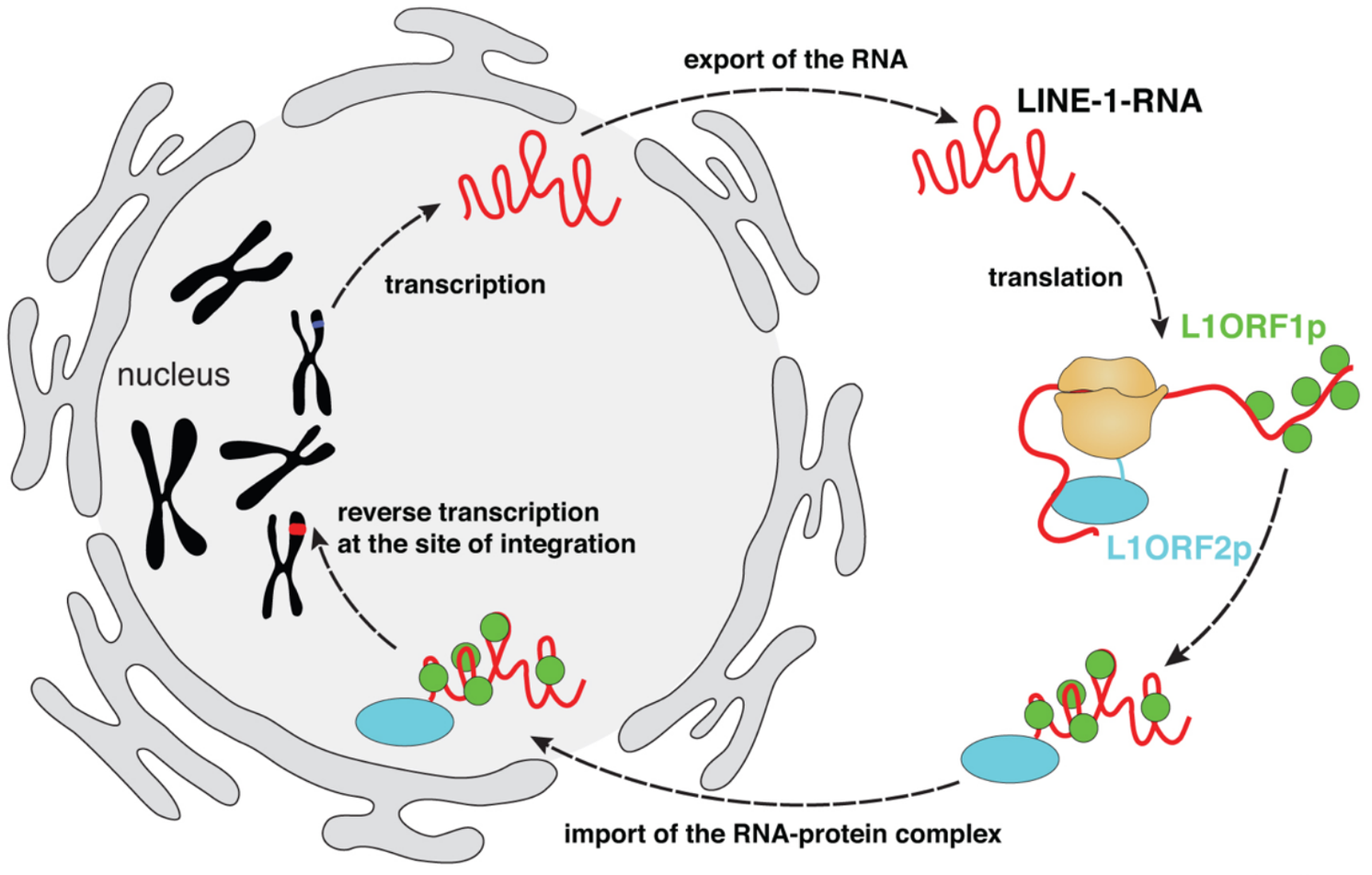
- Introduction: TEs in the human genome
- Transposable elements activity in healthy tissues
- Scientific computing on Seven Bridges
- Future steps
    - Transposable elements activity in human tumors

Agenda

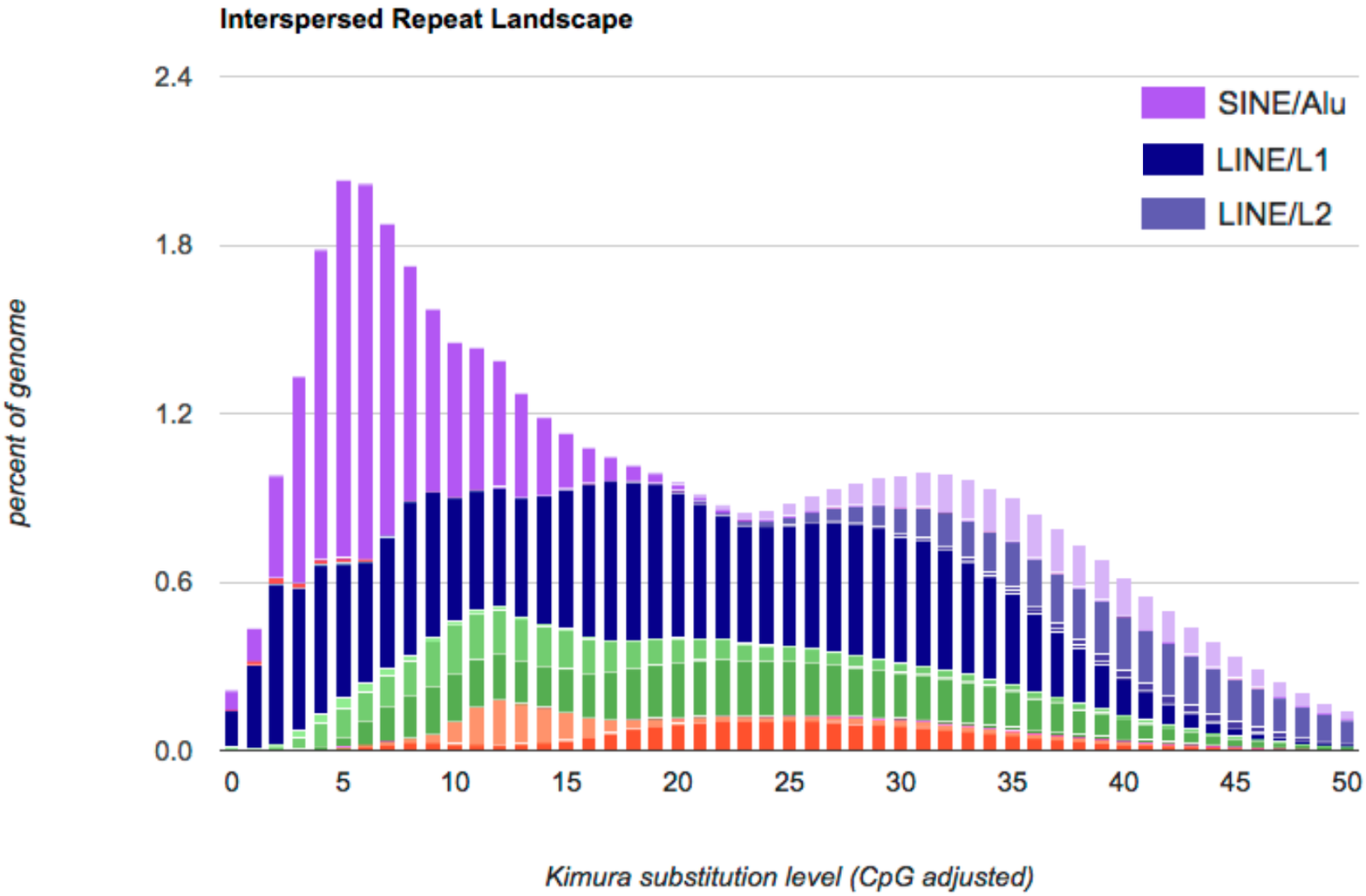
- **Introduction: TEs in the human genome**
- Transposable elements activity in healthy tissues
- Scientific computing on Seven Bridges
- Future steps
  - Transposable elements activity in human tumors

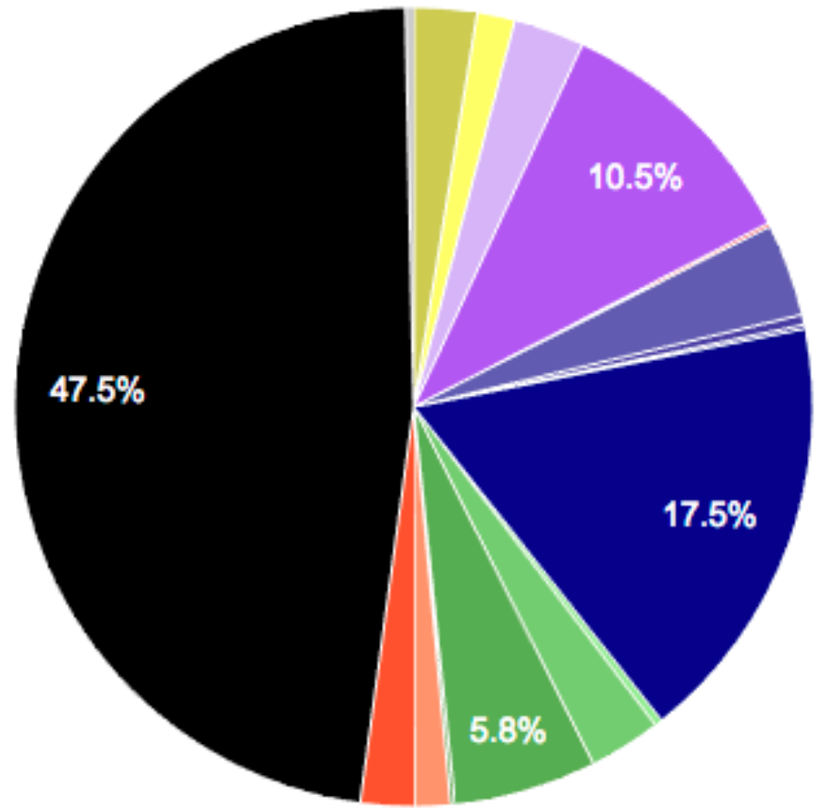
(retro)Transposable elements





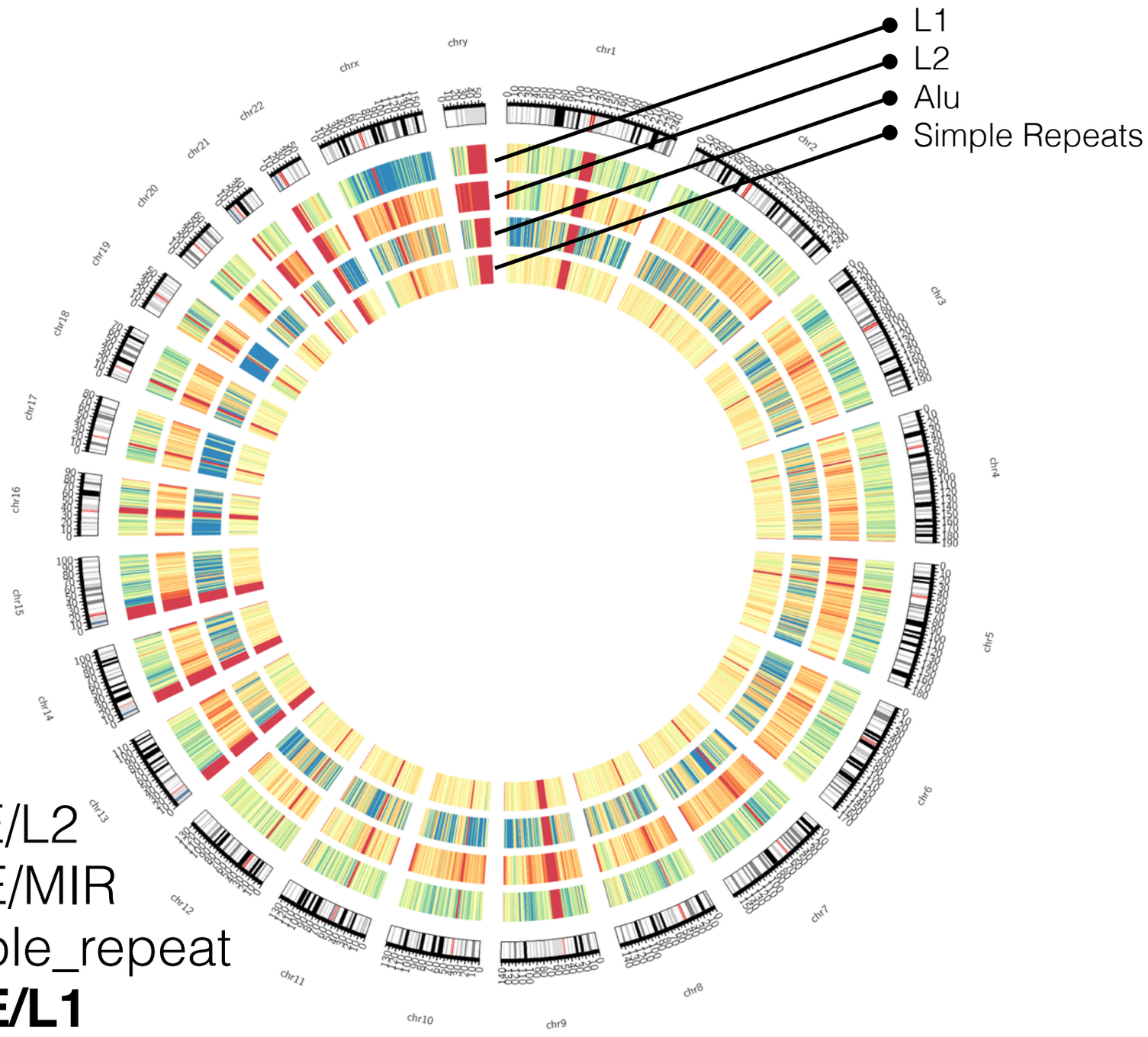
Repetitions in our genome





534,592	LINE/L2
611,412	SINE/MIR
700,455	Simple_repeat
978,523	**LINE/L1**
1,238,490	**SINE/Alu**

Transposable element density



534,592	LINE/L2
611,412	SINE/MIR
700,455	Simple_repeat
978,523	**LINE/L1**
1,238,490	**SINE/Alu**

Some open questions about LINE-1 biology:

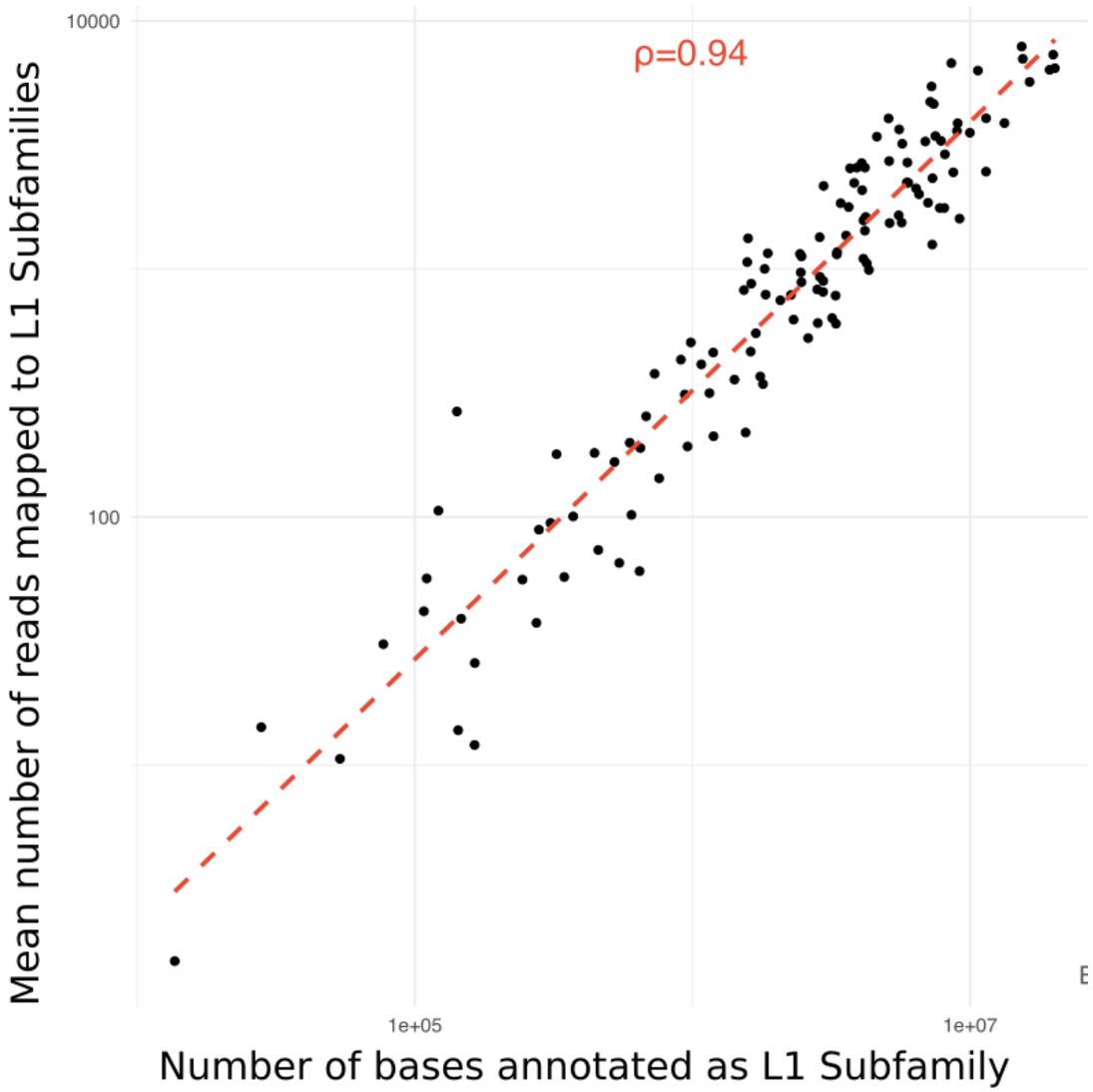
- Are LINE-1 completely silenced in somatic tissue?
- There are papers describing the activity of LINE-1 in tumors, but surprisingly, none about ALUs
- LINE-1 causes huge disruptions in the genome, what is their impact in tumor evolution?

Agenda

- Introduction: TEs in the human genome
- **Transposable elements activity in healthy tissues**
- Scientific computing on Seven Bridges
- Future steps
  - Transposable elements activity in human tumors

TeXP model

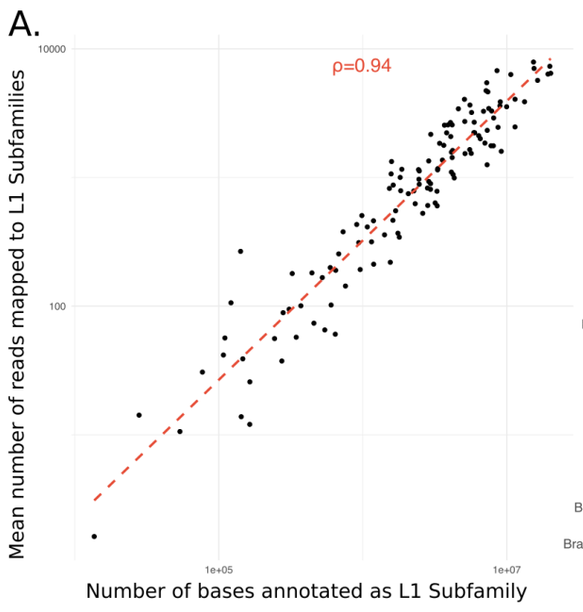
- Are LINE-1 completely silenced in somatic tissue?

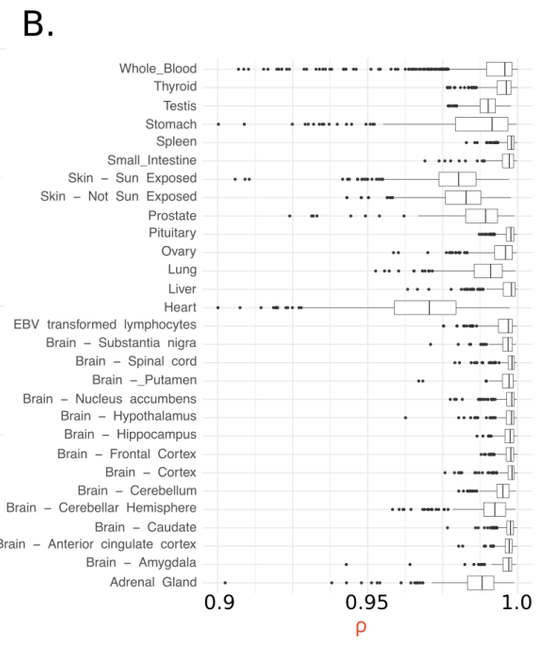


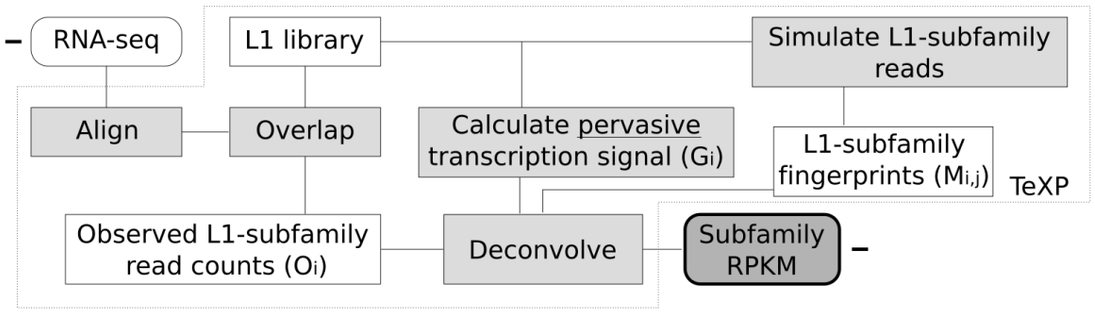
TeXP model

$$O_i = T(G_i \epsilon_{pervasive} + M_{i,j} \varepsilon_j)$$

Noise from pervasive transcription







Oi is the observed number of reads mapping to L1Hs;

T is the total number of reads mapped to L1 instances;

Gi defines the proportion of L1 bases in the genome annotated as subfamily i;

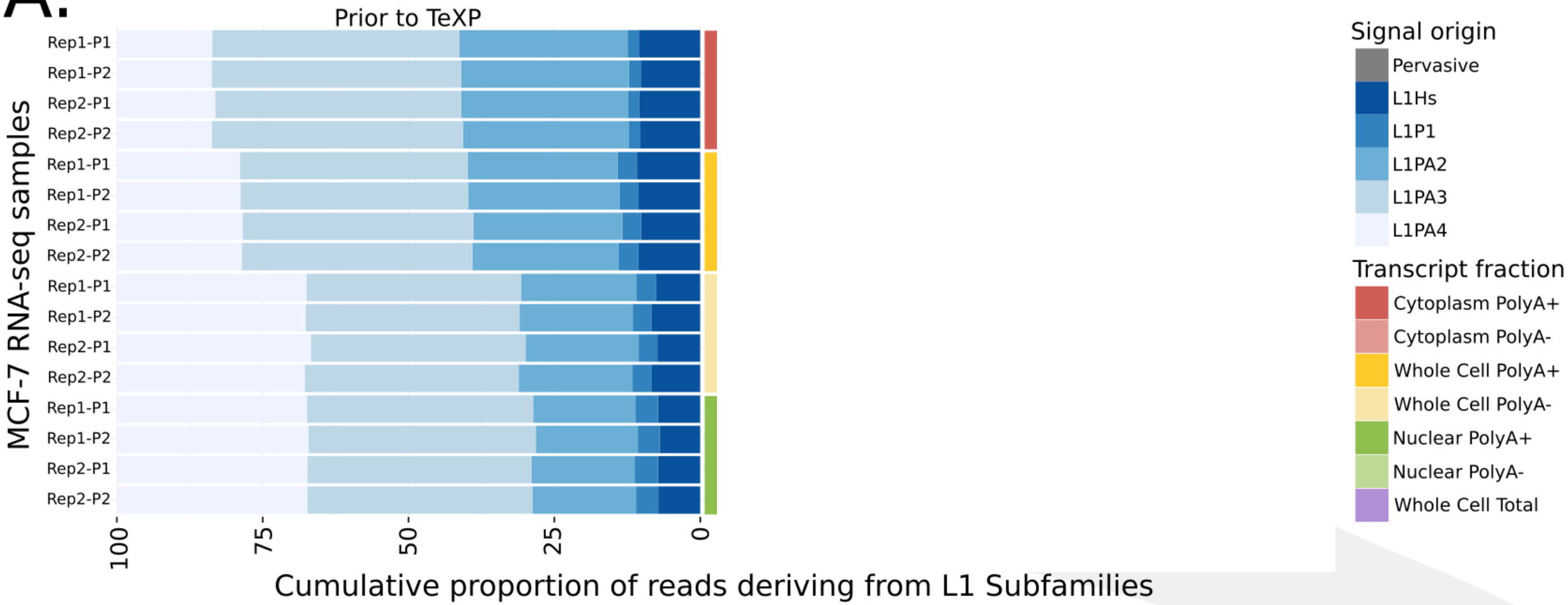
Epervasive is the percentage of reads emanating from pervasive transcription;

Mi,j is the mappability fingerprint (defined bellow) that describes what is the proportion of reads emanating from the signal i that maps to L1 subfamily j;

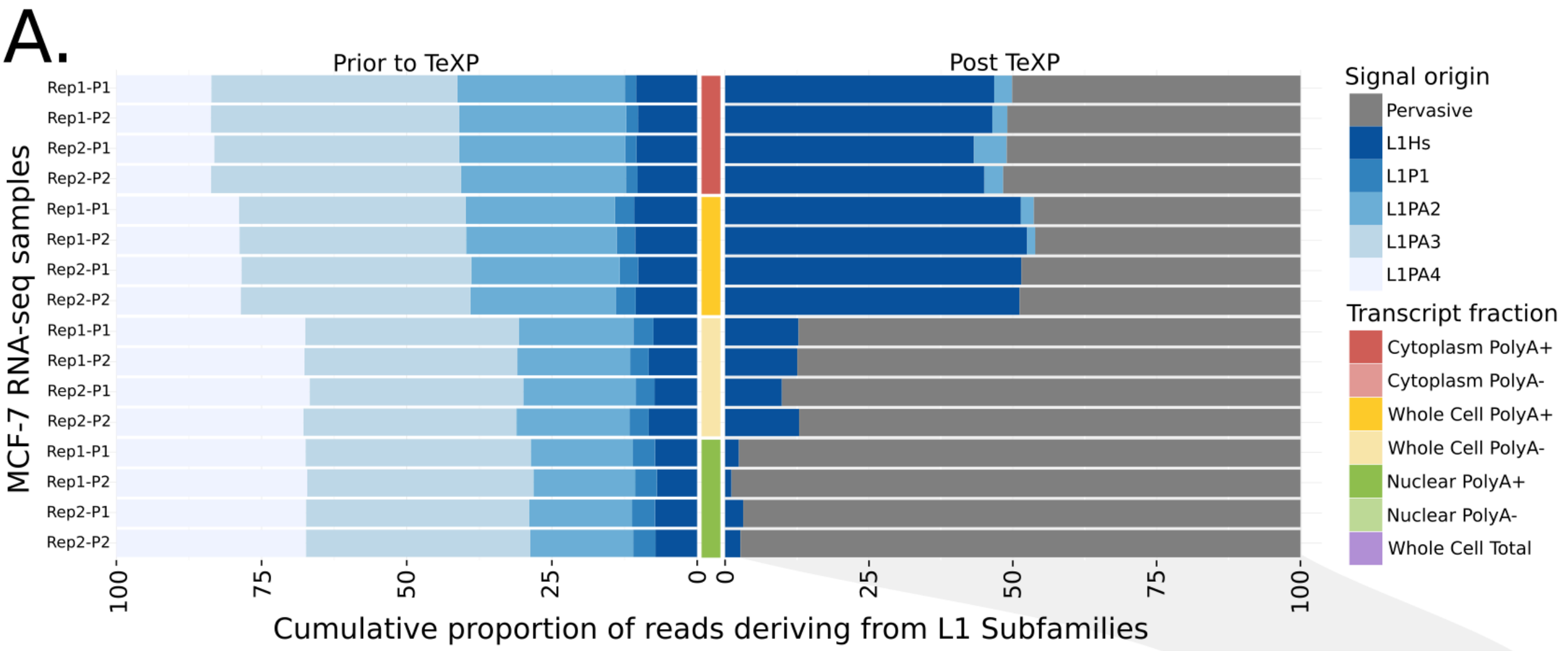
Ej is the percentage of reads emanating from the L1 Subfamily .

TeXP Validation

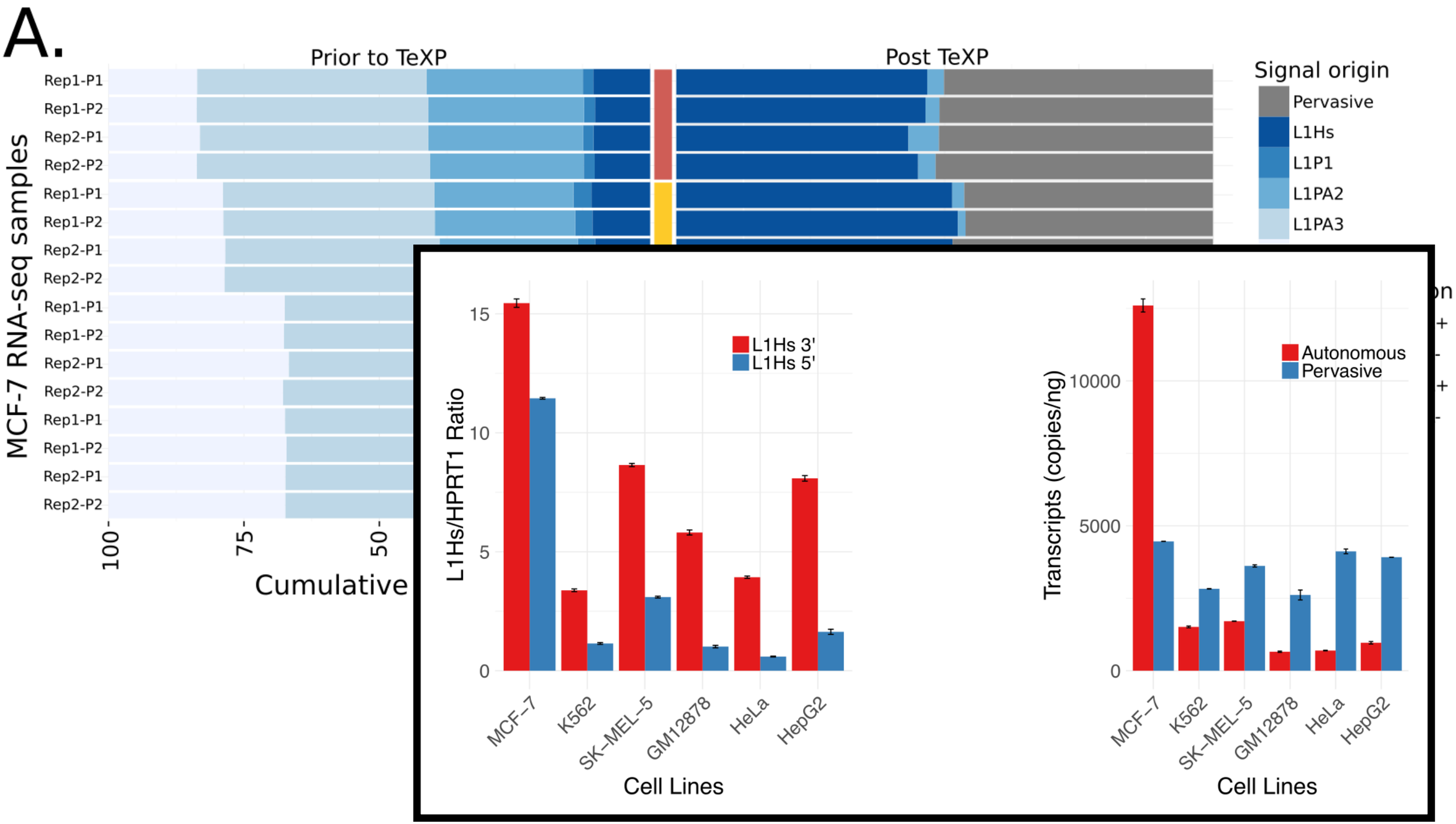
A.



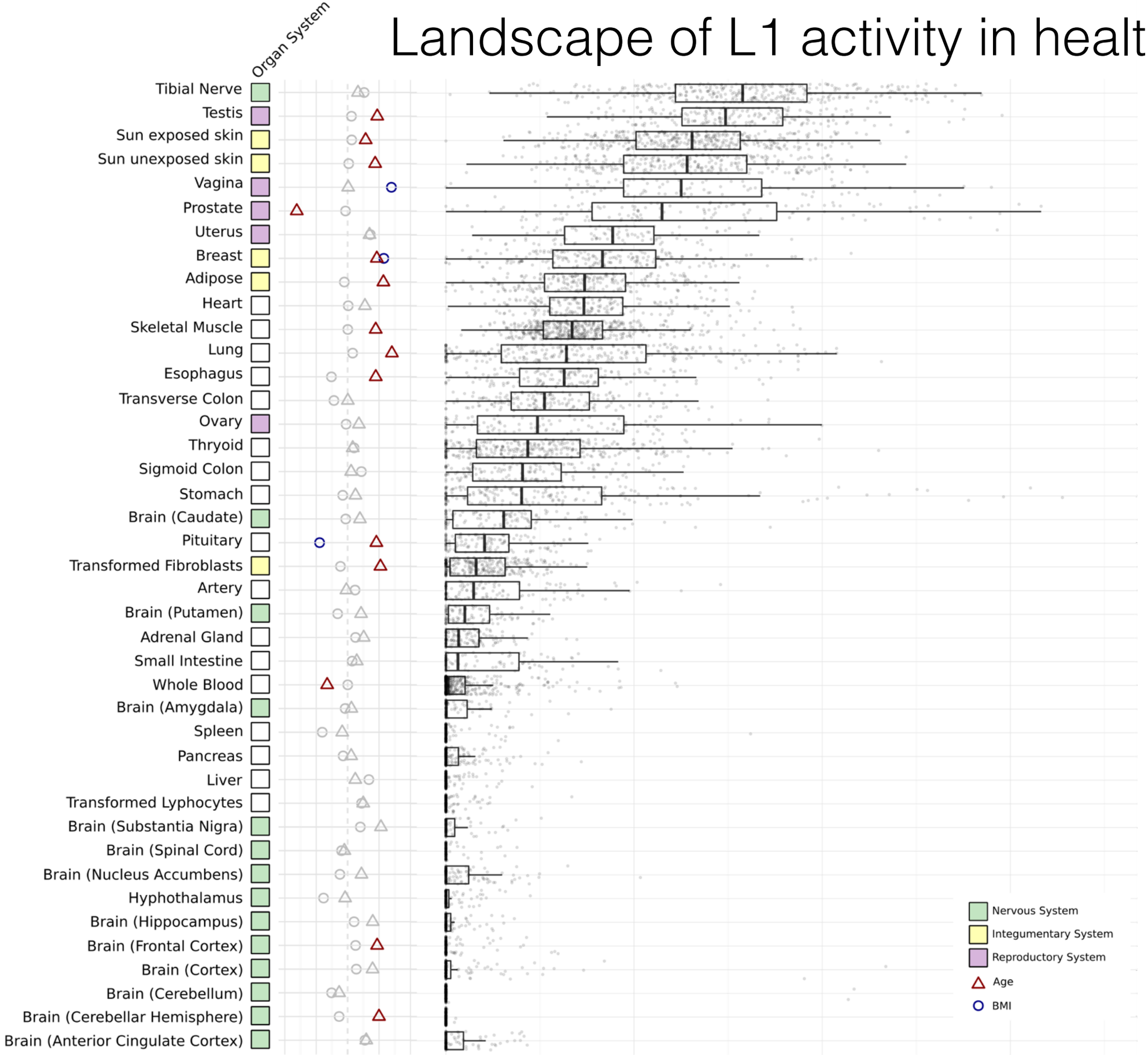
TeXP Validation



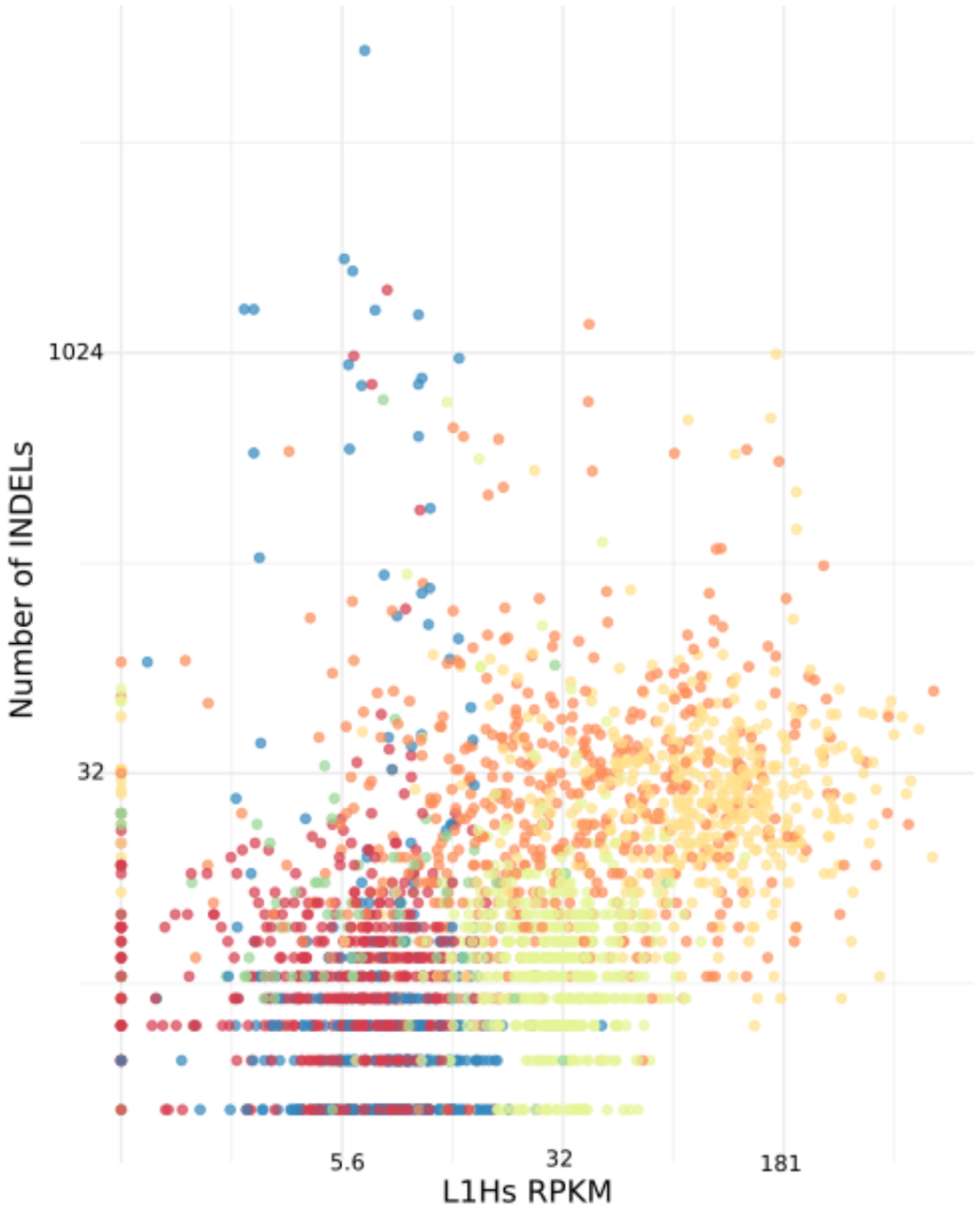
TeXP Validation

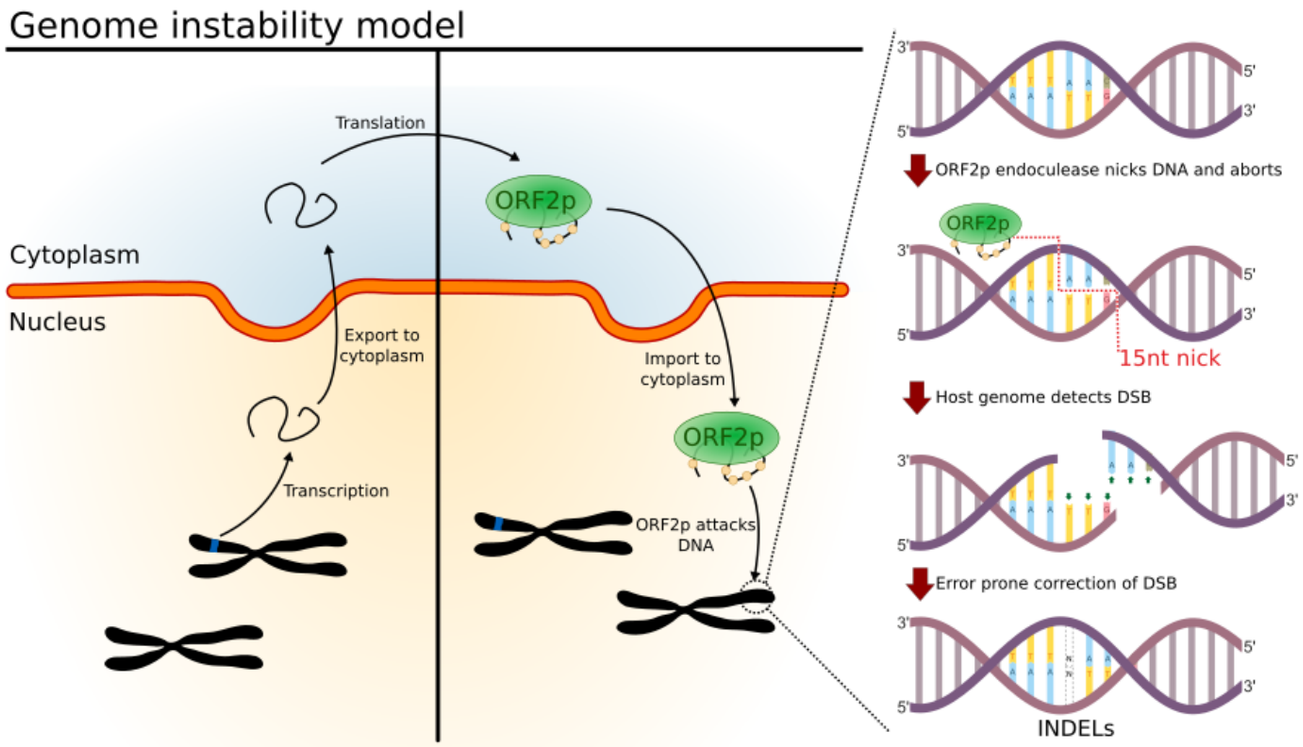


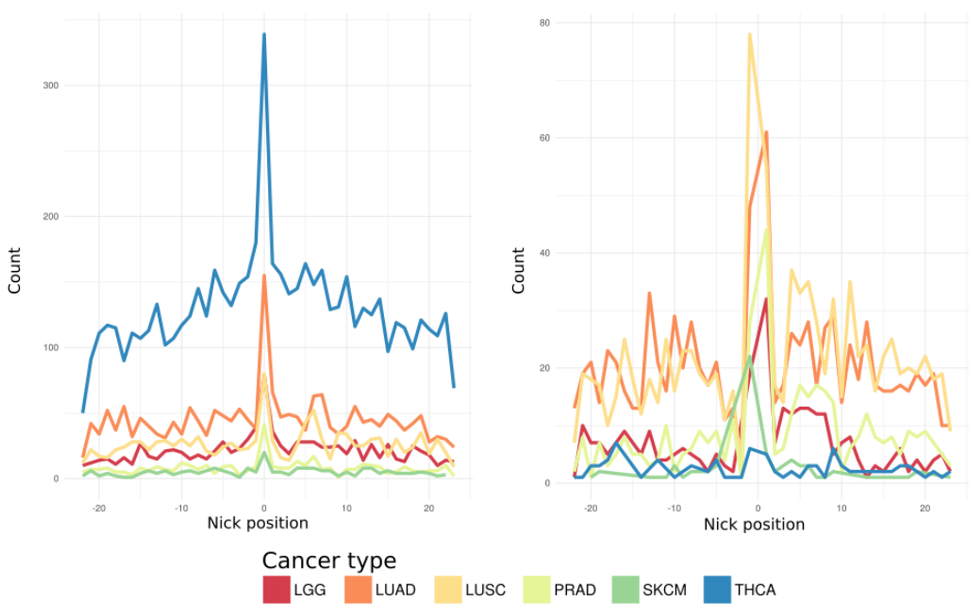
Landscape of L1 activity in healthy tissues



Correlation between L1 expression and Number of INDELs in tumors





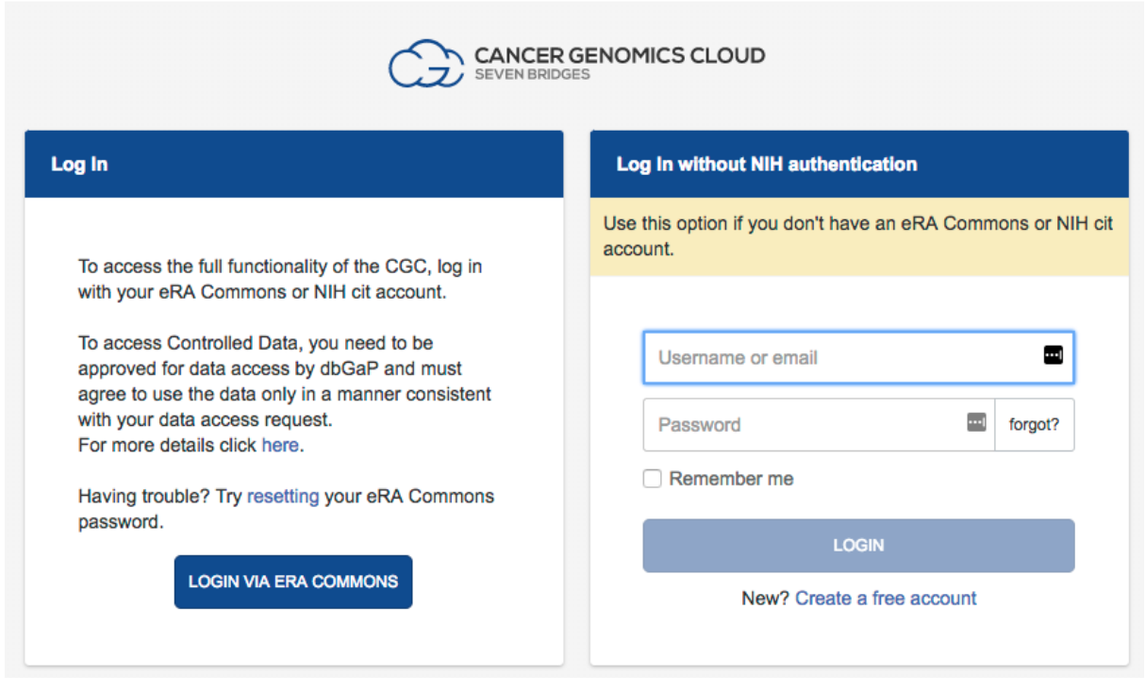


Agenda

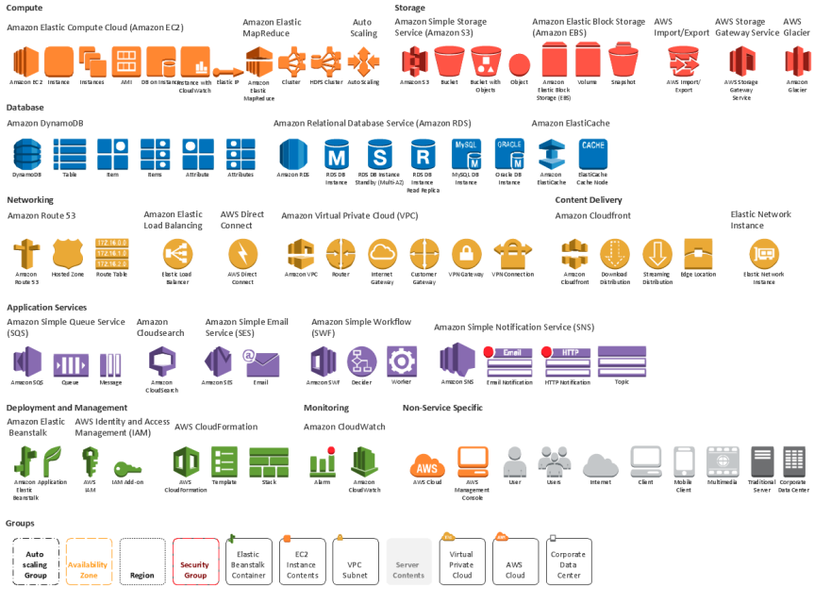
- Introduction: TEs in the human genome
- Transposable elements activity in healthy tissues
- **Scientific computing on Seven Bridges**
- Future steps
    - Transposable elements activity in human tumors

Seven Bridges CGC

Front-End

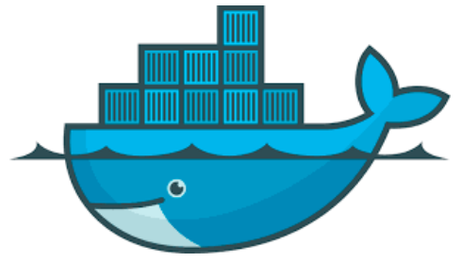


Amazon AWS



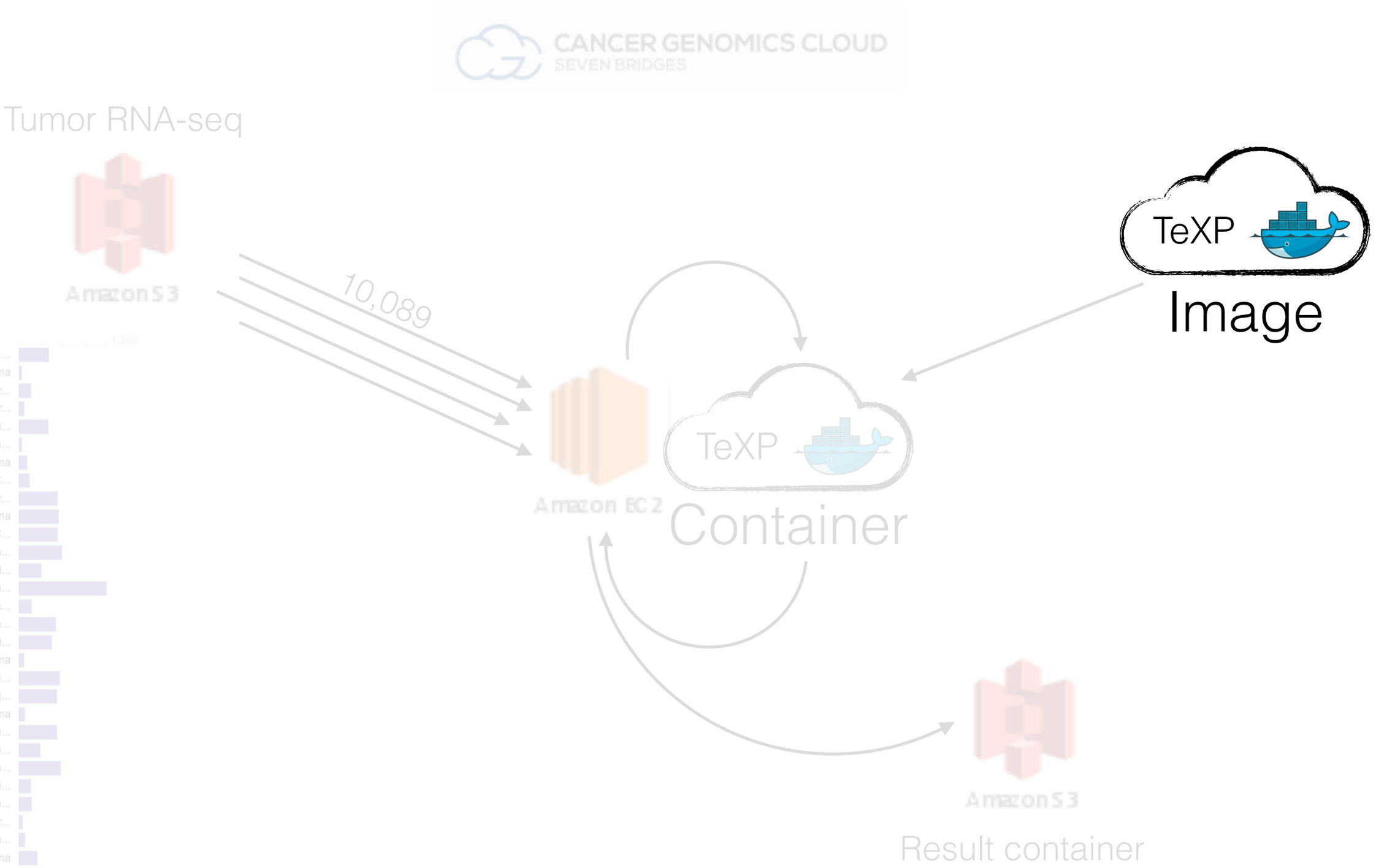
Rabix CWL

[Reproducible Analysis for Bioinformatics]

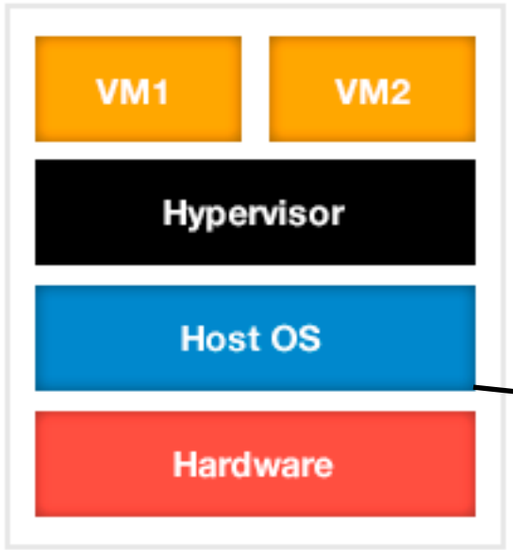


Docker

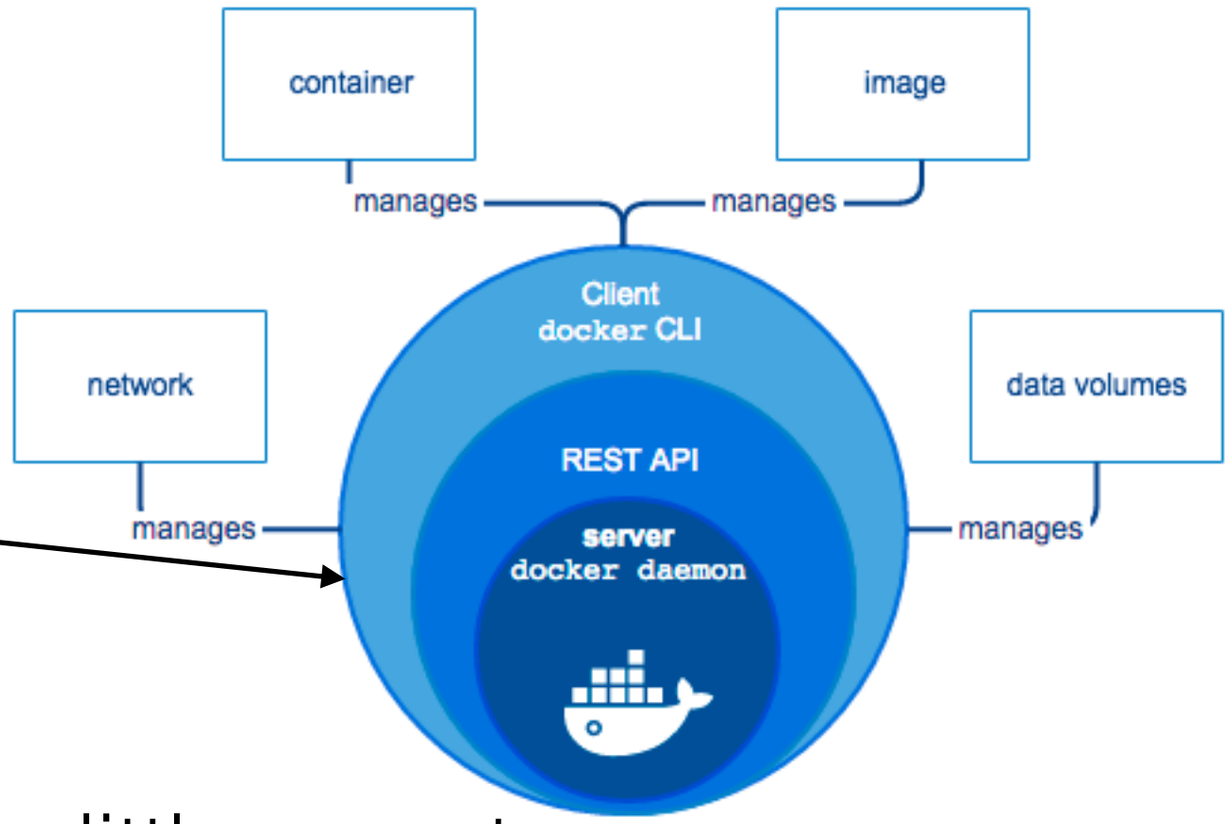
CGC Life cycle



Docker?



Hosted Hypervisor



except for a dirty little secret:
MacOS and Windows don’t have this especial container implementation. So Both run over VMs!

Defining your docker "VM"

Dockerfile

```
FROM debian:stretch
RUN apt-get update

##################
#Install binaries dependencies
##################

RUN apt-get install -y \
    bedtools=2.26.0+dfsg-3 \
```

$ docker build -t fnavarro/temp:latest ./Dockerfile

Defining your docker "VM"

Dockerfile

```
FROM debian:stretch
RUN apt-get update

##################
#Install binaries dependencies
##################

RUN apt-get install -y \
    bedtools=2.26.0+dfsg-3 \
```

$ docker build -t fnavarro/temp:latest ./Dockerfile

Defining your docker "VM"

Dockerfile

```
FROM debian:stretch
RUN apt-get update

##################
#Install binaries dependencies
##################

RUN apt-get install -y \
   bedtools=2.26.0+dfsg-3 \
```

$ docker build -t fnavarro/temp:latest ./Dockerfile

Defining your docker "VM"

Dockerfile

```
FROM debian:stretch
RUN apt-get update

##################
#Install binaries dependencies
##################

RUN apt-get install -y \
    bedtools=2.26.0+dfsg-3 \
```

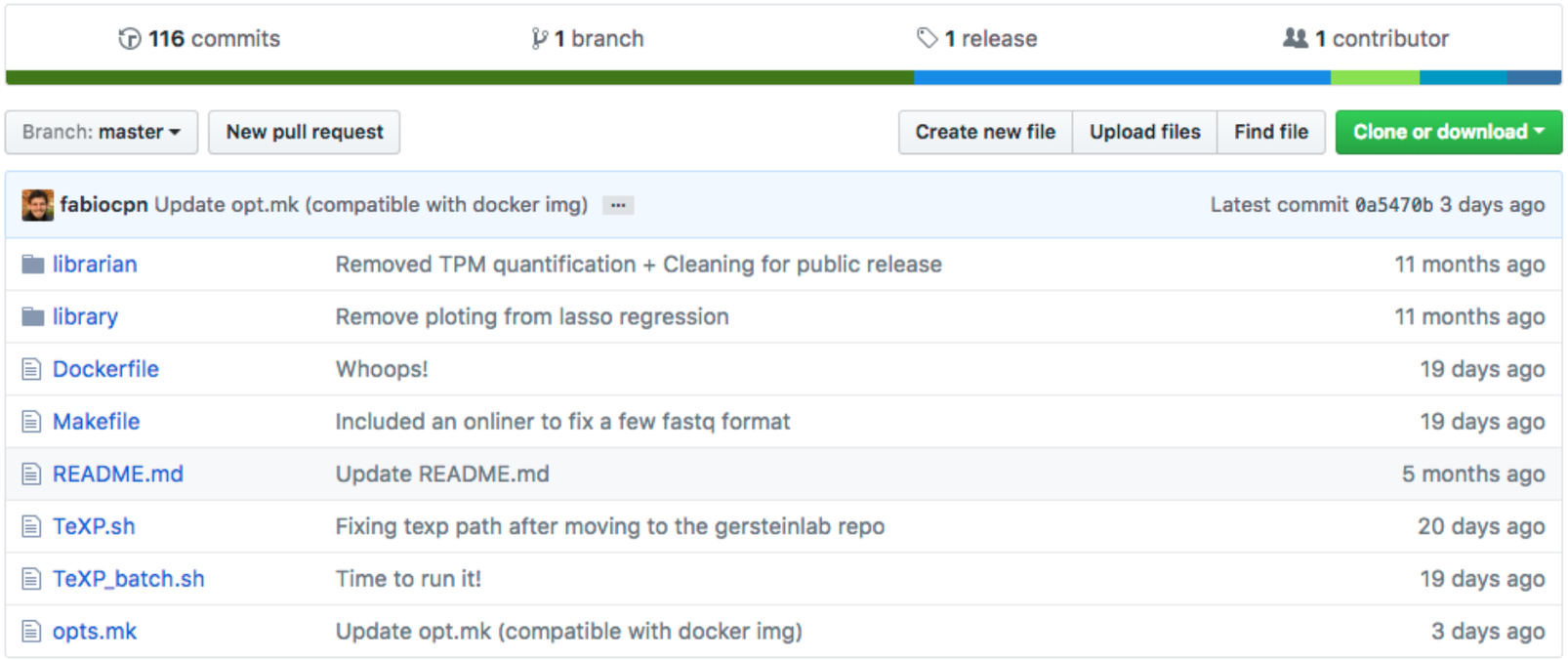
$ docker build -t fnavarro/temp:latest ./Dockerfile

```
Building in Docker Cloud's infrastructure...
Cloning into '.'...
.
Reset branch 'master'
Your branch is up-to-date with 'origin/master'.
Pulling cache layers for fnavarro/texp:latest...
Done!
KernelVersion: 4.4.0-93-generic
Arch: amd64
BuildTime: 2017-08-17T22:50:04.828747906+00:00
ApiVersion: 1.30
Version: 17.06.1-ce
MinAPIVersion: 1.12
GitCommit: 874a737
Os: linux
GoVersion: go1.8.3
Starting build of index.docker.io/fnavarro/texp...
Step 1/12 : FROM debian:stretch
---> da653cee0545
Step 2/12 : RUN apt-get update
---> Using cache
---> 17904f776045
Step 3/12 : RUN apt-get install -y bedtools
---> Using cache
---> eea5e6affff4
```

```
FROM debian:stretch
RUN apt-get update

#################
#Install binaries dependencies

RUN apt-get install -y \
        samtools=1.3.1-3 \
             […]
        wget

#################
#Install Wgsim

#################
#Download Libraries

RUN wget rep_annotation.hg38.tar.bz2
RUN wget bowtie2.hg38.tar.bz2

#################
#Install R packages dependencies

#################
#Install TeXP

ADD https://api.github.com/repos/gersteinlab/texp/git/refs/heads/master version.json
RUN mkdir -p /src; \
        cd /src ; \
        git clone https://github.com/gersteinlab/texp.git

WORKDIR /src/texp/
CMD ["/src/texp/TeXP.sh"]
```

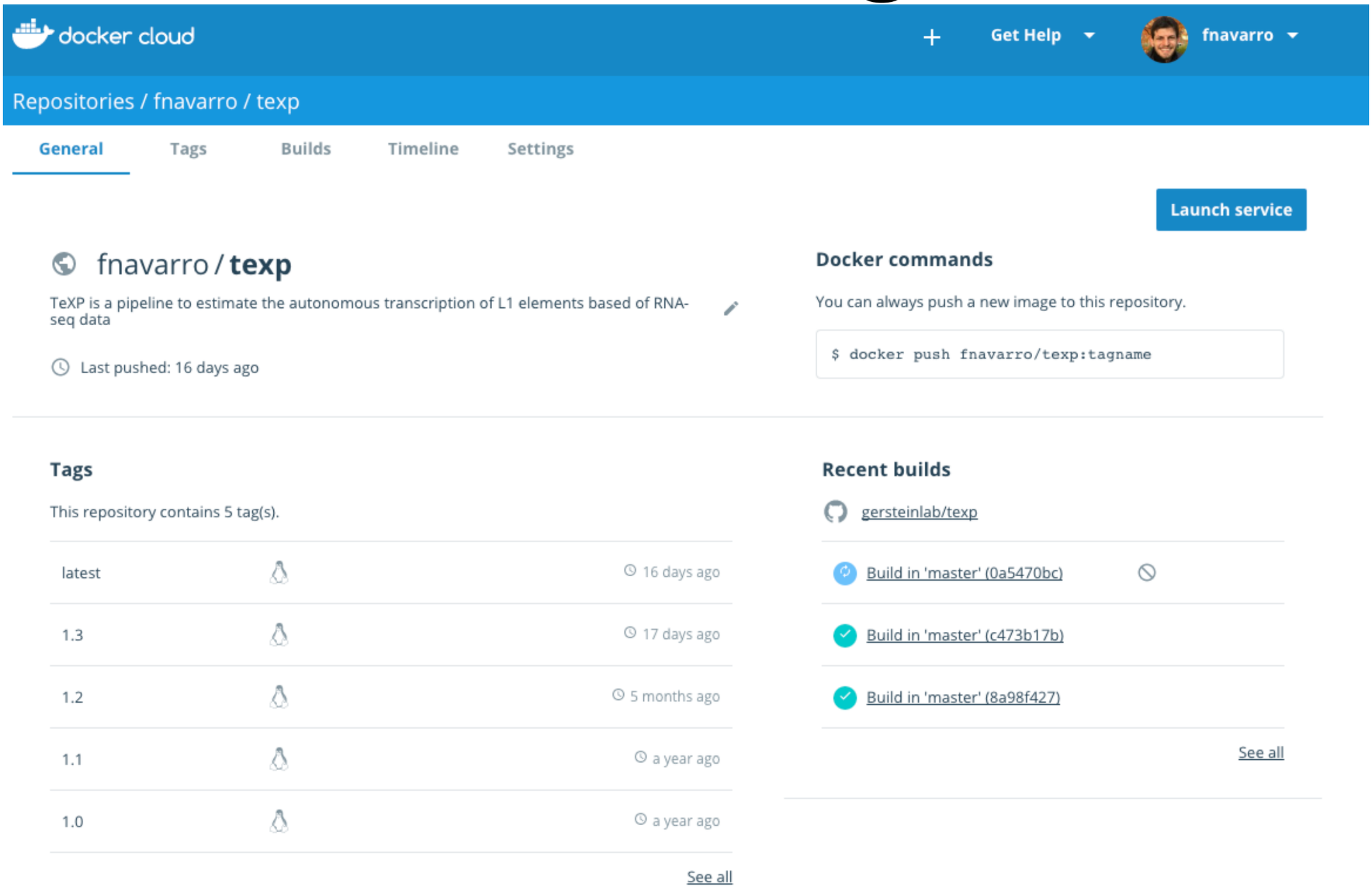
TeXP Dockerfile

```
FROM debian:stretch
RUN apt-get update

#################
#Install binaries dependencies

RUN apt-get install -y \
        samtools=1.3.1-3 \
              […]
        wget

#################
#Install Wgsim

#################
#Download Libraries

RUN wget rep_annotation.hg38.tar.bz2
RUN wget bowtie2.hg38.tar.bz2

#################
#Install R packages dependencies

#################
#Install TeXP

ADD https://api.github.com/repos/gersteinlab/texp/git/refs/heads/master version.json
RUN mkdir -p /src; \
        cd /src ; \
        git clone https://github.com/gersteinlab/texp.git

WORKDIR /src/texp/
CMD ["/src/texp/TeXP.sh"]
```



https://github.com/gersteinlab/texp.git

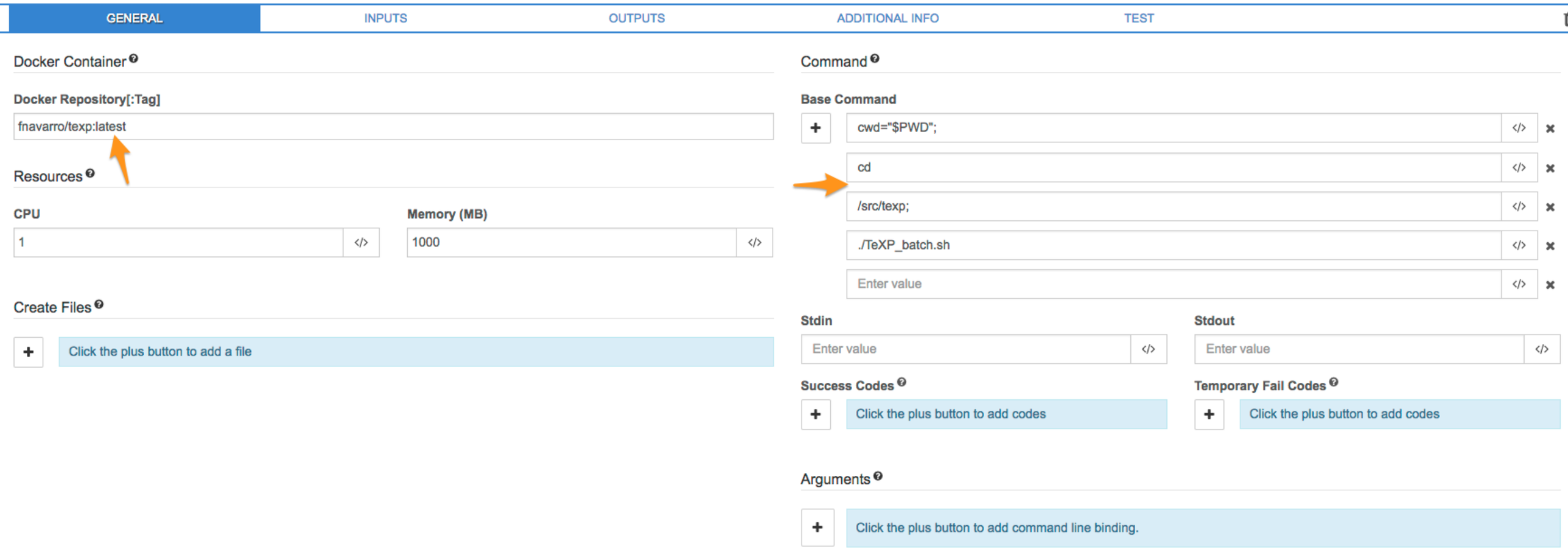
docker.io + github



docker run -it fnavarro/texp:latest /bin/bash

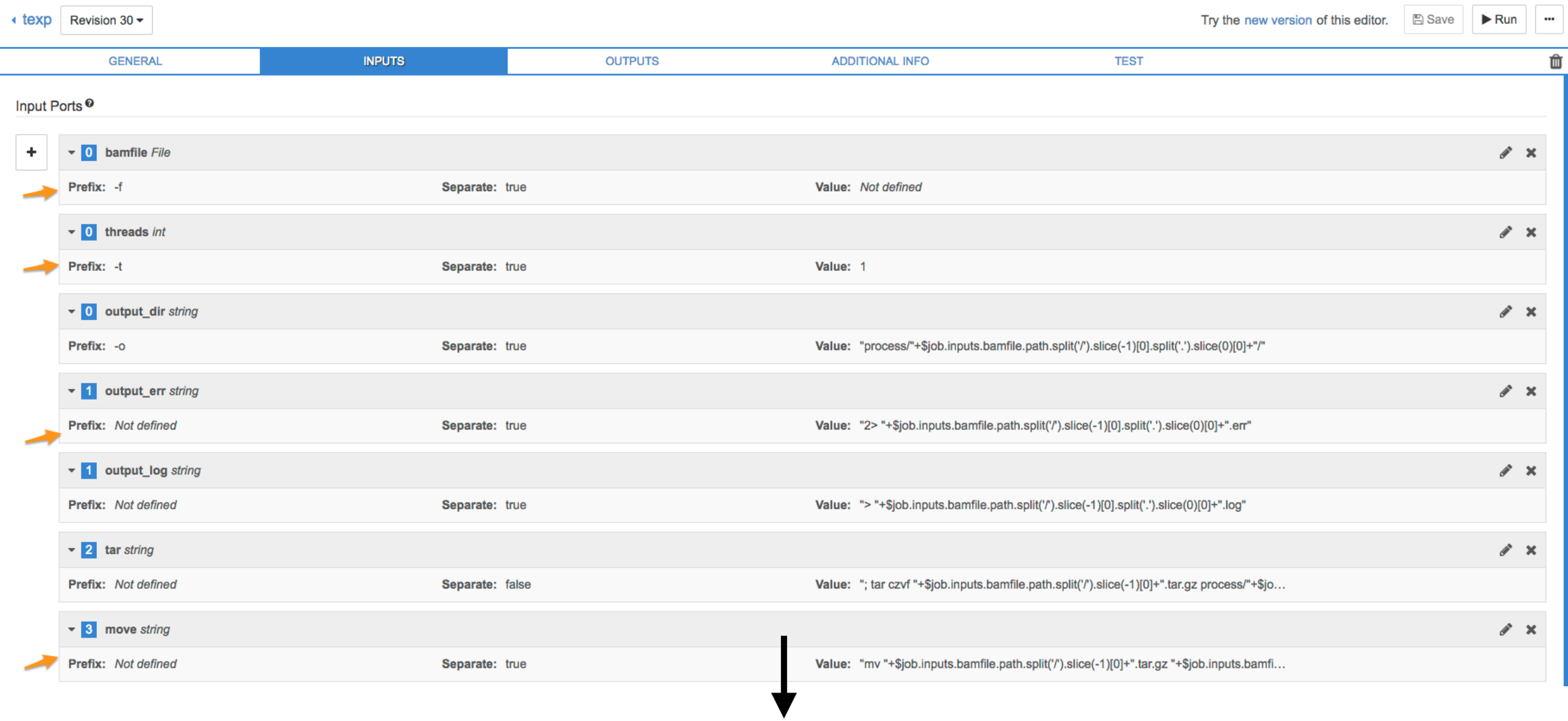
Setting up TeXP on CGC

Setting up TeXP on CGC (1)



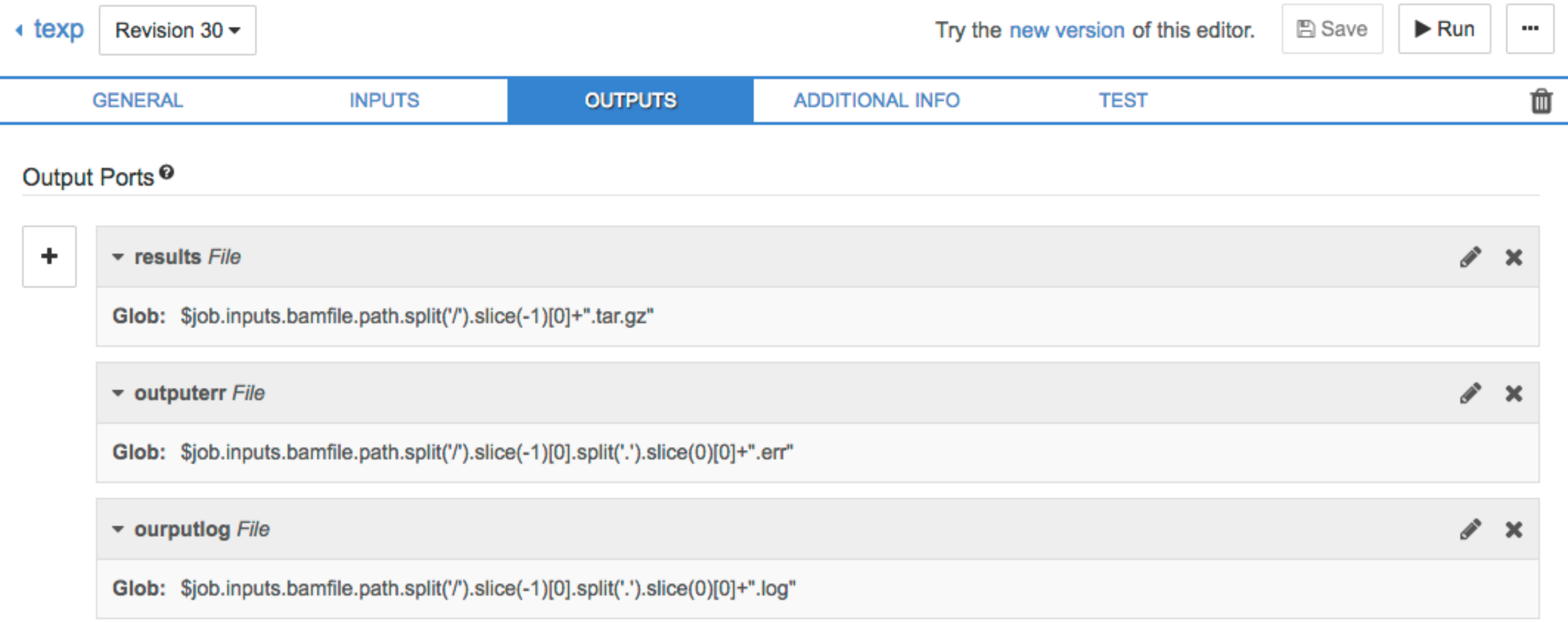
docker pull fnavarro/texp:latest; cwd="$PWD"; cd /src/texp; ./TeXP_batch.sh

Setting up TeXP on CGC (2)



-f /path/to/bamfile.ext -t 1 -o process/bamfile/ 2> bamfile.err > bamfile.log ; tar czvf bamfile.ext.tar.gz process/bamfile/; mv bamfile.ext.tar.gz bamfile.log bamfile.err $cwd

Setting up TeXP on CGC (3)

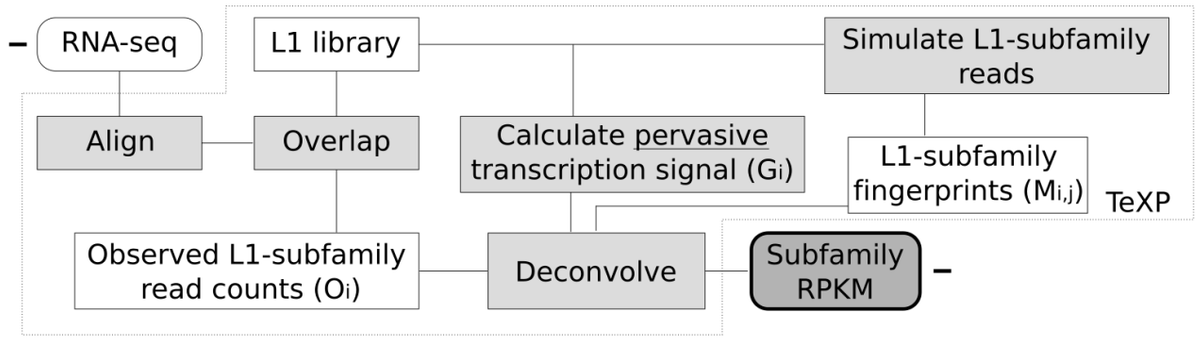


Setting up TeXP on CGC (4)

Create App workflow

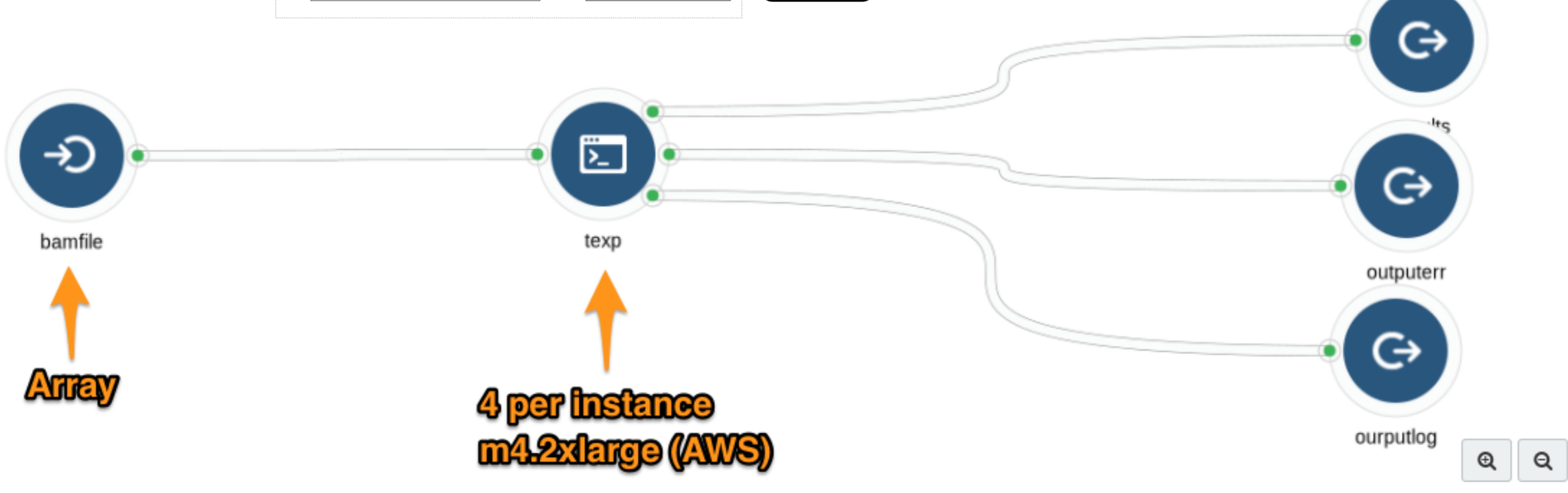
Description

No description.



process/UNCID_2332590/130718_UNC12-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.err
process/UNCID_2332590/130718_UNC12-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.log
process/UNCID_2332590/130718_UNC12-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.stats

process/UNCID_2332590/130718_UNC12-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq:
L1HS_hg38.count
L1HS_hg38.count.corrected
L1HS_hg38.count.rpkm
L1HS_hg38.count.rpkm.corrected
L1HS_hg38.count.signal_proportions
qualityEncoding
read_length
re.bam
re.filtered.bam
re.filtered.bed
sorted.bam.bai
sorted.bam.tot
L1HS_hg38.signatures.txt



CGC Demo (3min)

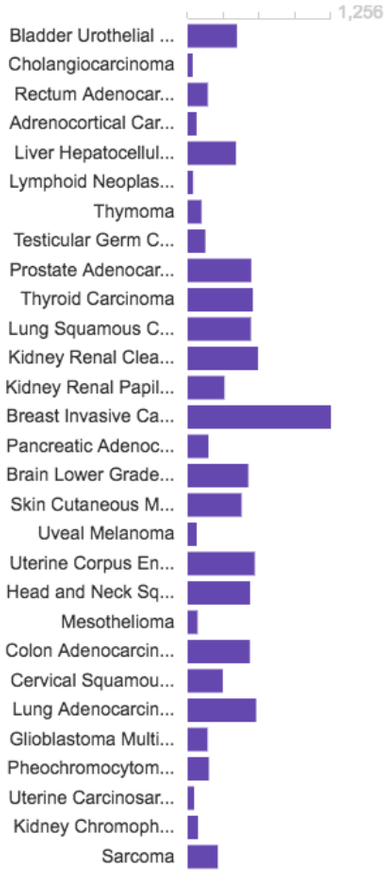
http://cgc.sbgenomics.com

CGC Life cycle

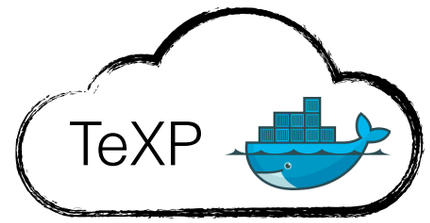
CANCER GENOMICS CLOUD
SEVEN BRIDGES

Tumor RNA-seq

Amazon S3



10,089



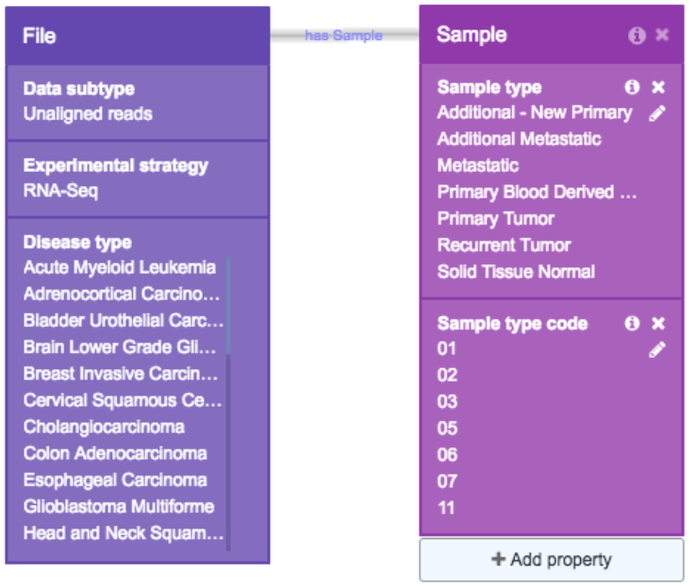
Image

Amazon EC2

Container

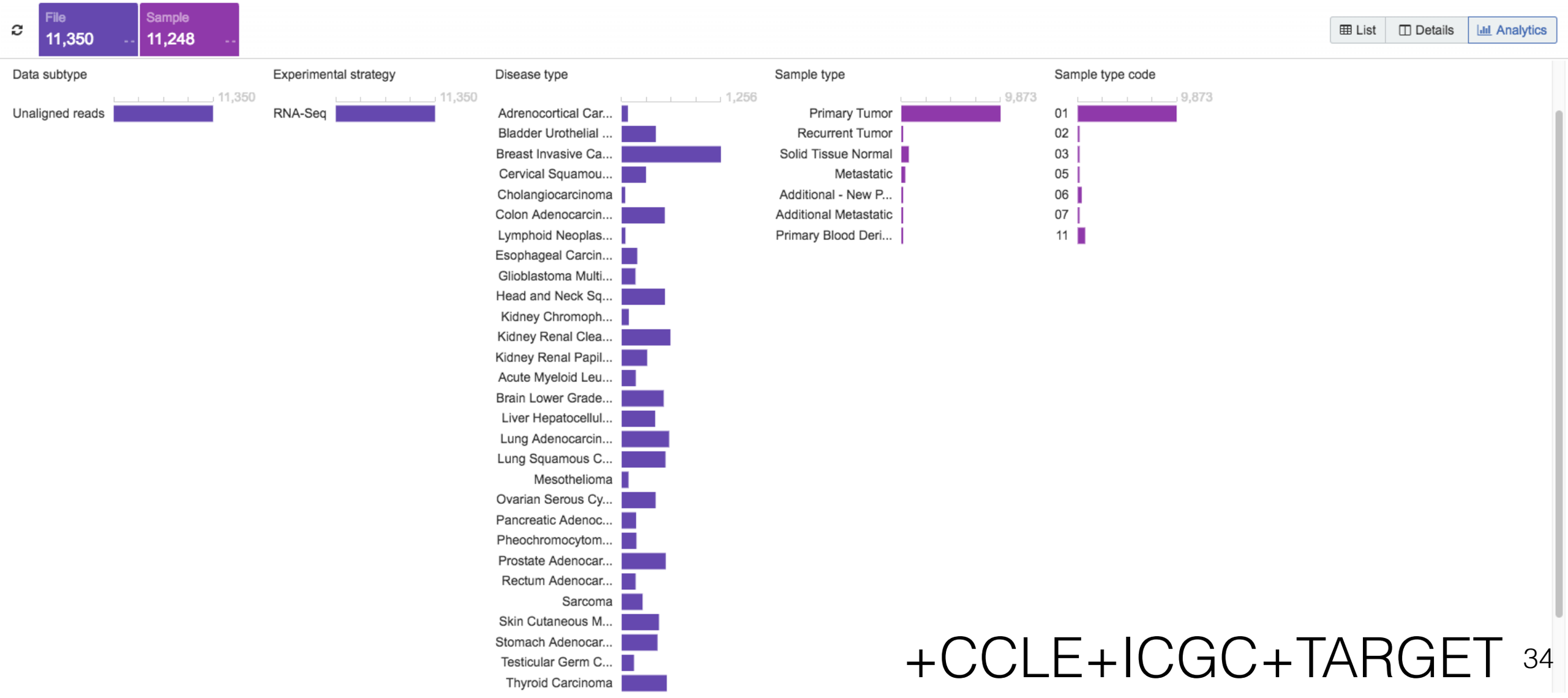
Amazon S3

Result container



TCGA data on CGC

(if you have authorized access)

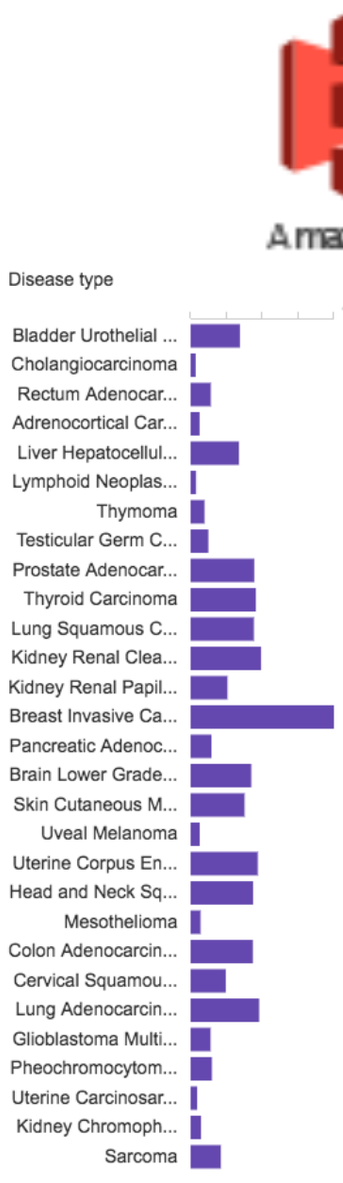


+CCLE+ICGC+TARGET

CGC Life cycle

CANCER GENOMICS CLOUD
SEVEN BRIDGES

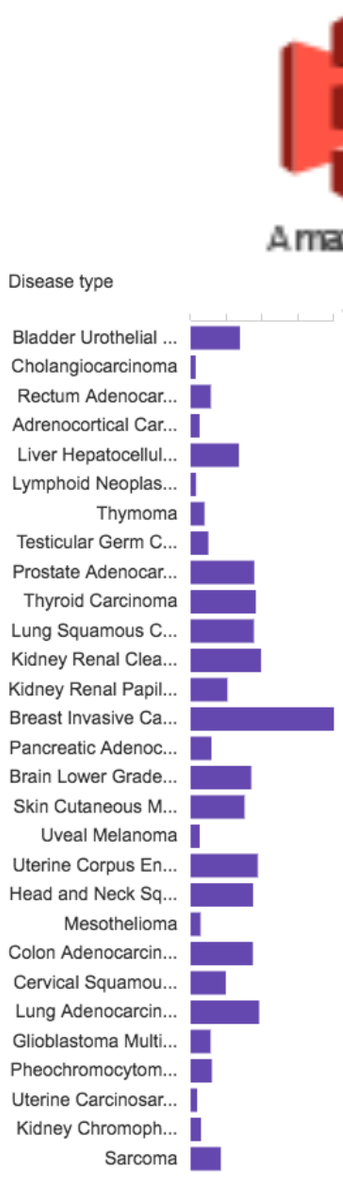
Tumor RNA-seq

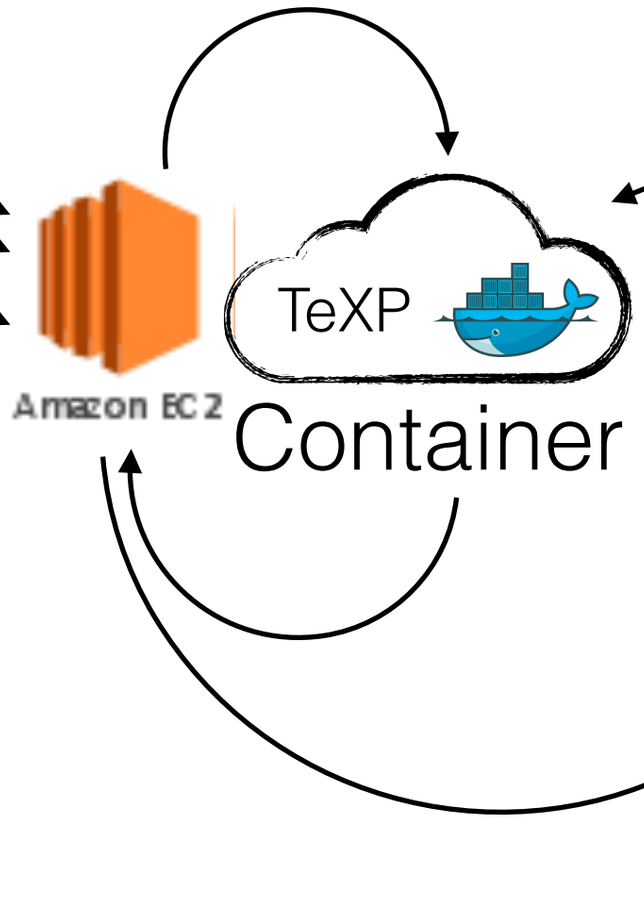


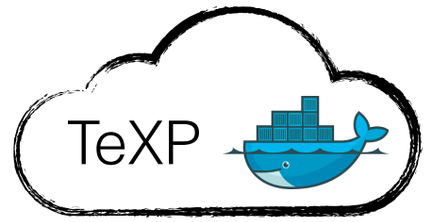
CGC Life cycle

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Tumor RNA-seq







80 instances
= 640 proc

Result container

Doings things programmatically

- Python API to manage launching;

```
my_app = find_app(api, my_project, app_name)
tar_files = find_files(api, my_project, “tar")
my_project = find_project(api, project_name)

my_task = api.tasks.create(name="texp run"+str(task_id),
                           project=my_project, app=my_app,
                           inputs=inputs, run=True)
```

- Python API to manage files (download/rm);

```
output_files = find_files(api, my_project, “breaks")
for i, curr_file in enumerate(output_files):
     curr_file.download(path=‘auto_'+curr_file.name)
     curr_file.delete()
```

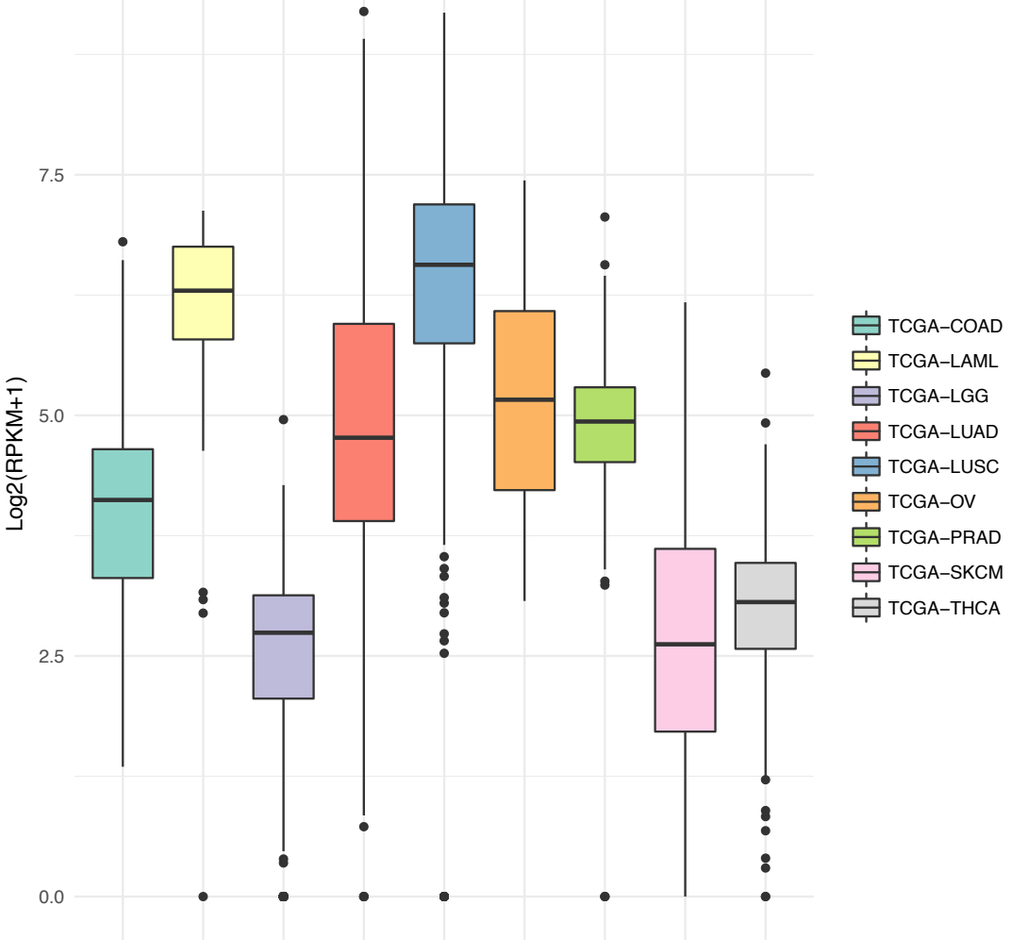
demo texp_6?

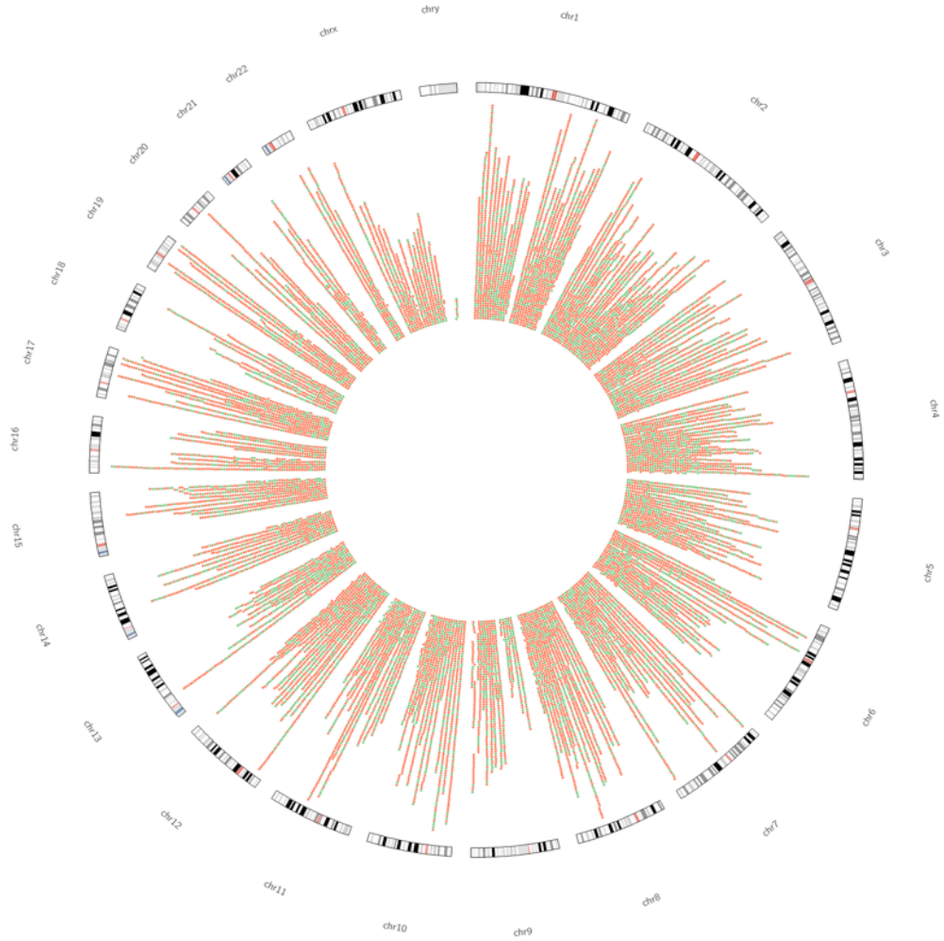
What have we learned so far?

- It’s fast-ish
- Mind the overhead of learning how to use all these tools
- It is kind of cheap
  - Break point analysis for all 4,880 TCGA WGS
    $1,800 - $0.36 p/ sample
  - TeXP for 3,900 RNA-seq samples
    $1200 - $0.30 p/ sample

Future steps

- Further explore the relationship between L1 transcription and genome instability





- Quantify the retrotransposition of ALUs in tumors

  75 samples -> 10,646 putative insertions

Acknowledgments

Mark Gerstein

Joel Rozowsky
Sushant Kumar
Prashant Emani
Timur Galeev
Gamze Gursoy
Jinrui Xu
Arif Harmanci
Eric Yu
Jacob Hoops

Charles Lee

Lauren Bellfy
Eliza Cerveira
Qihui Zhu
Chengsheng Zhang