

Transposable elements and Scientific cloud computing

Group Meeting - Jan 2018

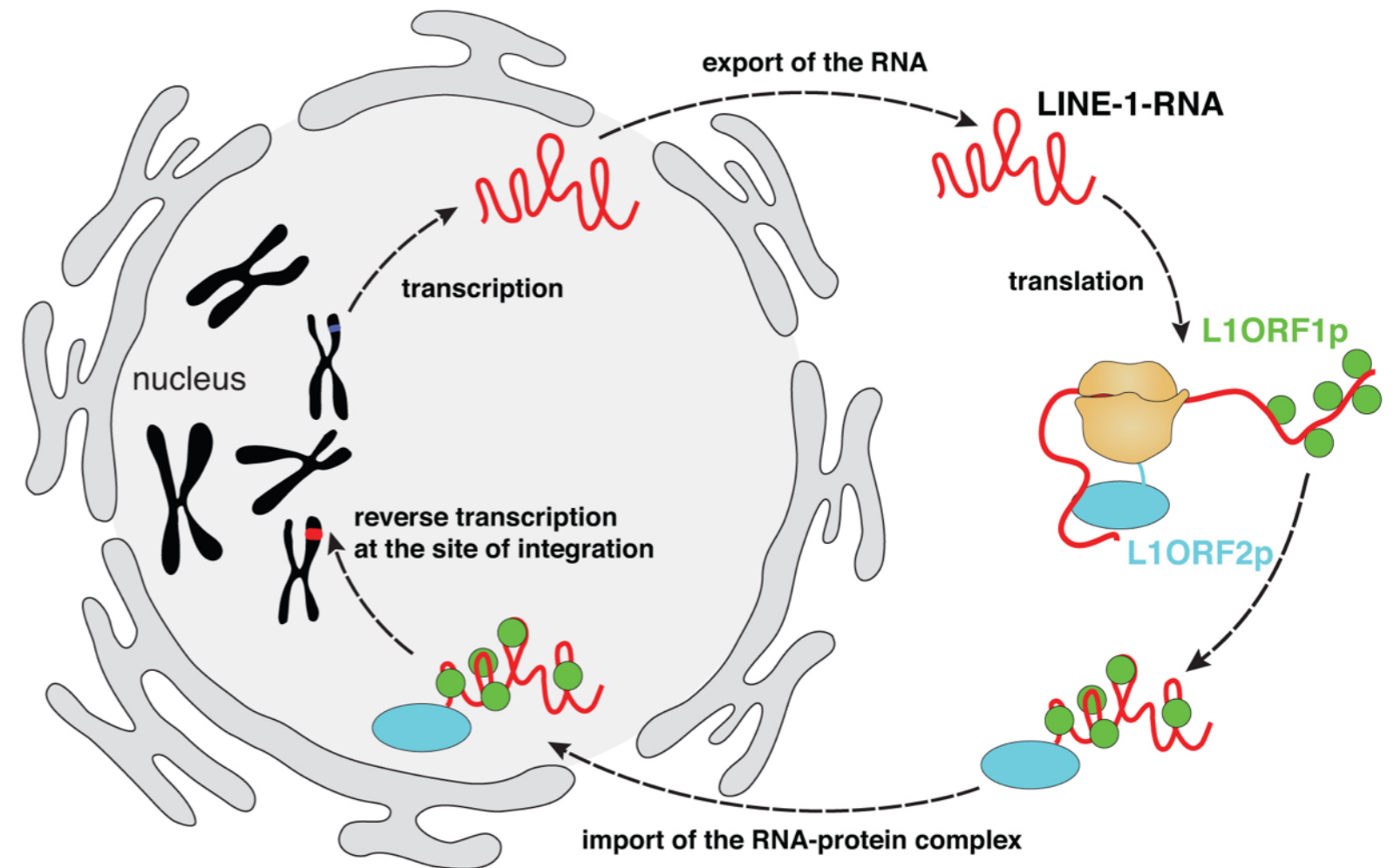
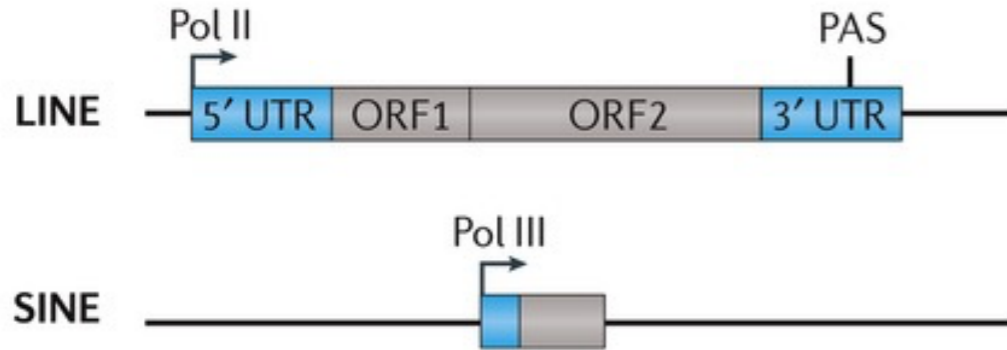
Agenda

- 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

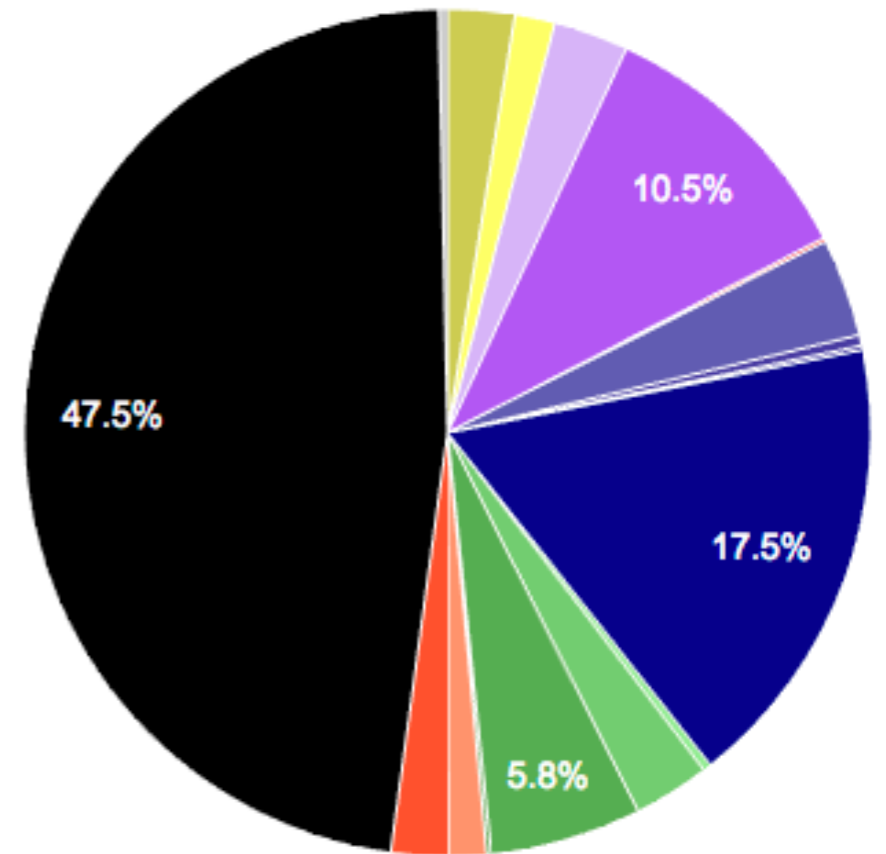
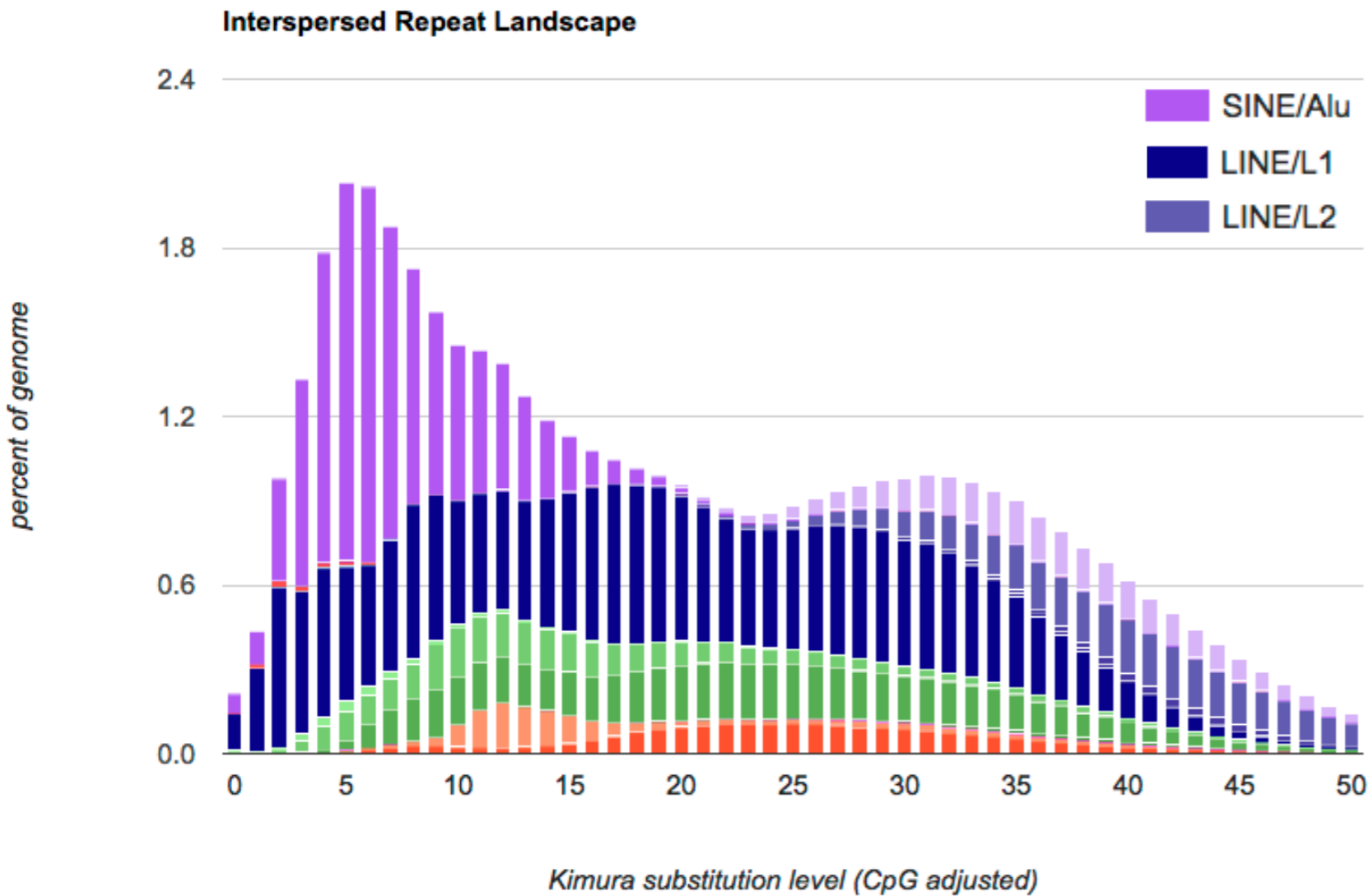
Agenda

- **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

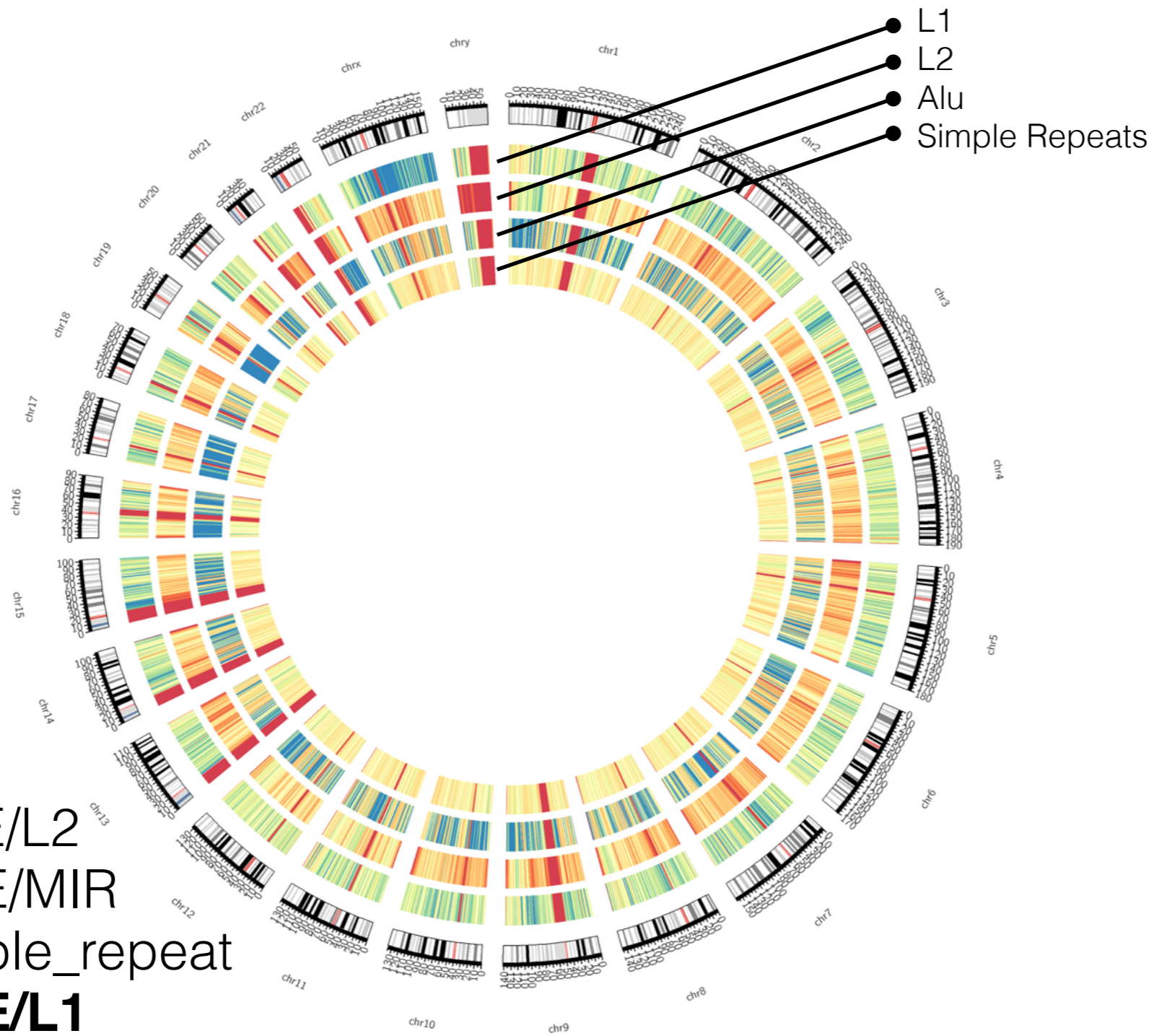


Repetitions in our genome



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
611,412
700,455
978,523
1,238,490

LINE/L2
SINE/MIR
Simple_repeat
LINE/L1
SINE/Alu

Some open questions about LINE-1 biology:

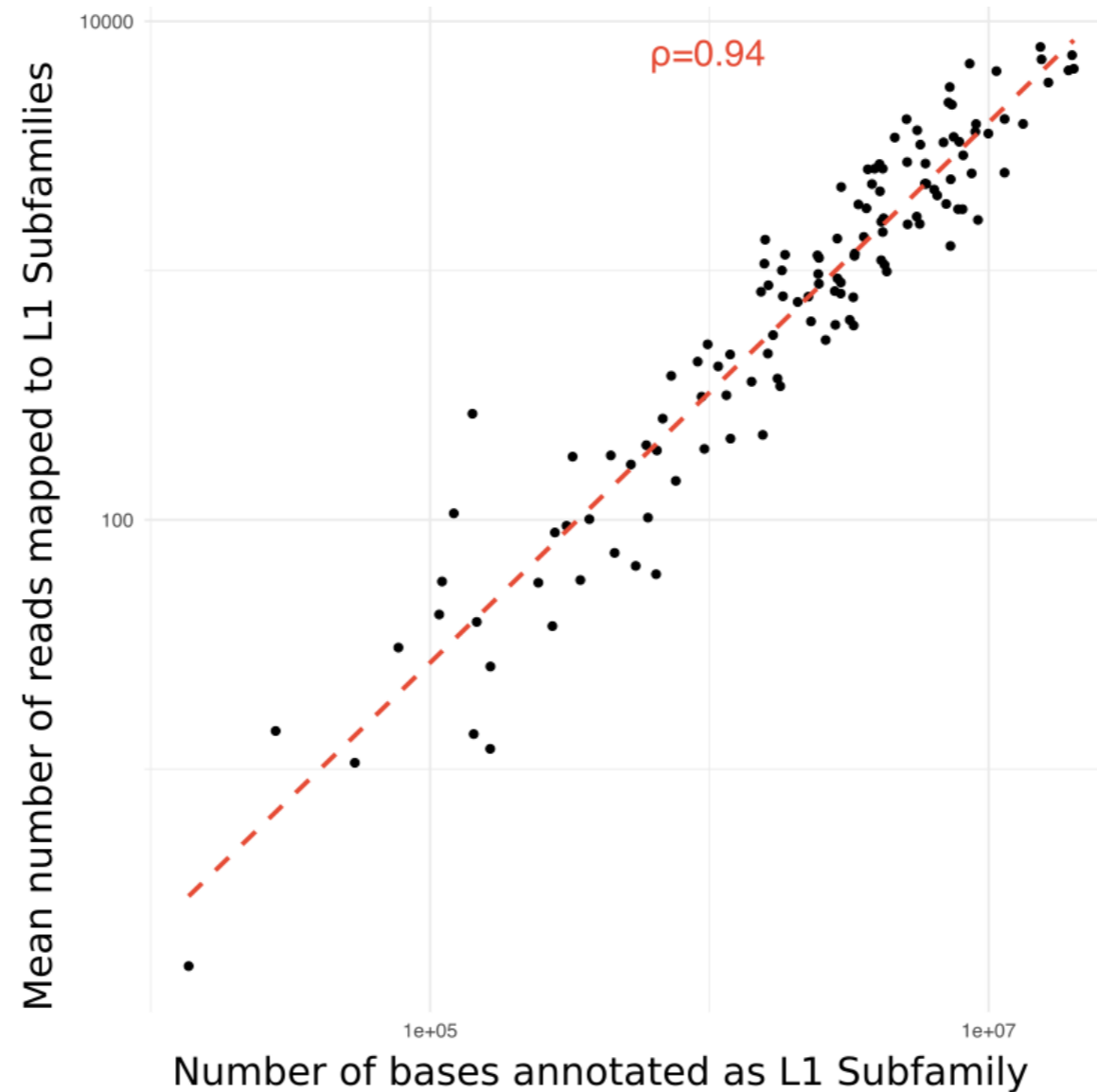
- Are LINE-1 completely silenced in somatic tissue?
- There are papers describing the activity of LINE-1 in tumors, but surprisingly, none about ALUs
- LINE-1 causes huge disruptions in the genome, what is their impact in tumor evolution?

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

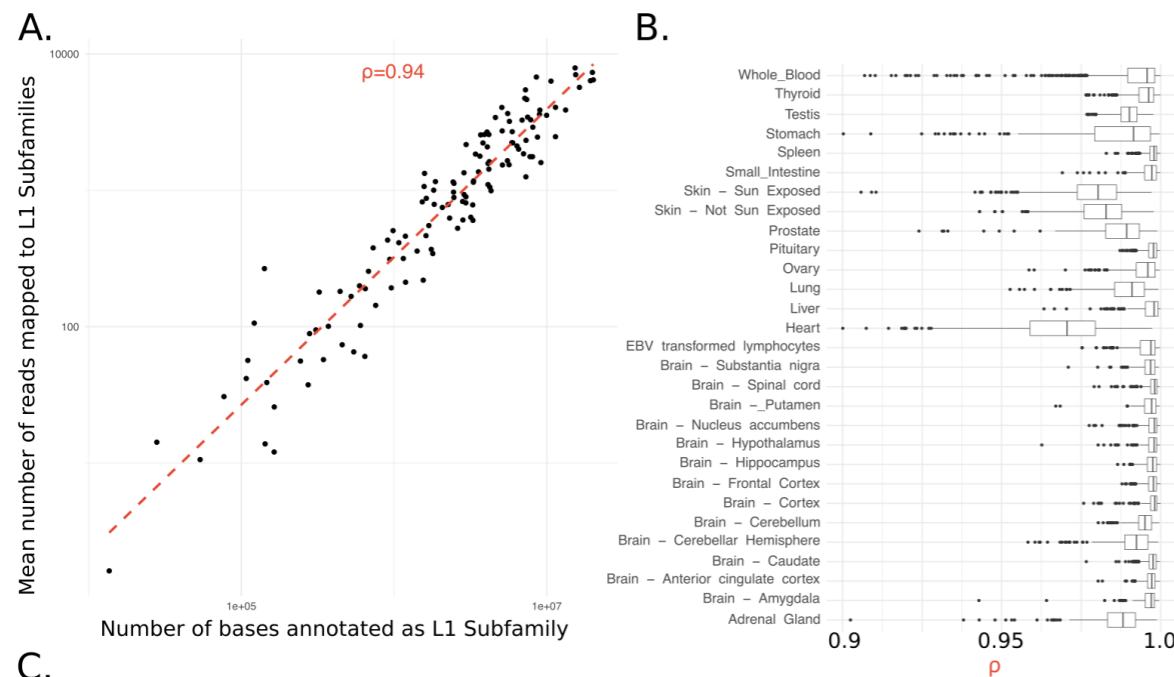
- Are LINE-1 completely silenced in somatic tissue?



TeXP model

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

Noise from pervasive transcription



O_i is the observed number of reads mapping to L1Hs;

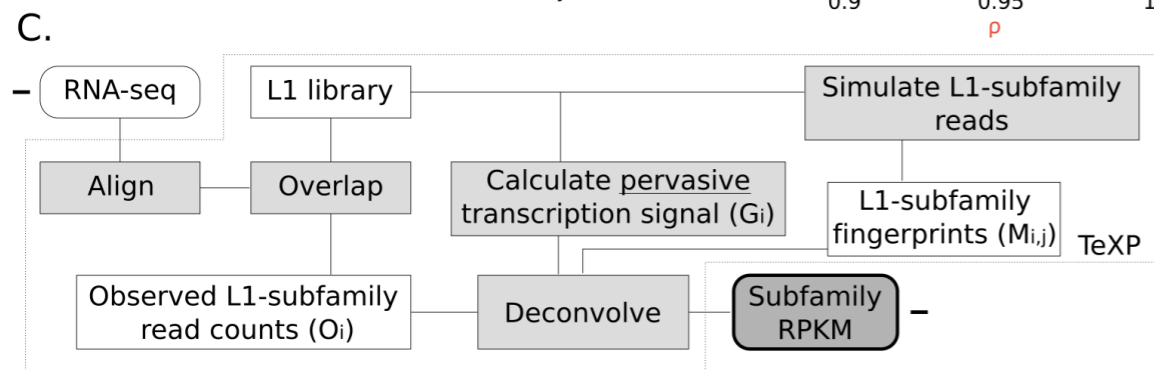
T is the total number of reads mapped to L1 instances;

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

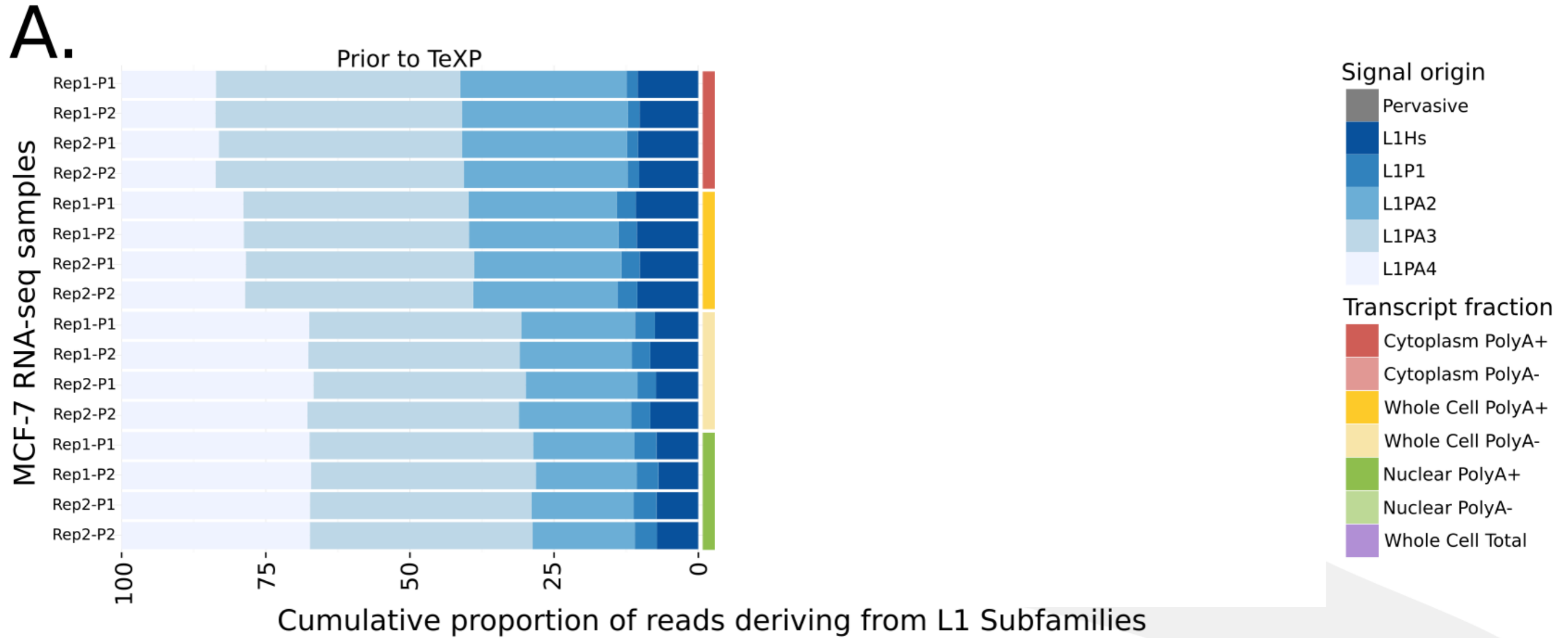
$\epsilon_{pervasive}$ is the percentage of reads emanating from pervasive transcription;

$M_{i,j}$ is the mappability fingerprint (defined below) that describes what is the proportion of reads emanating from the signal i that maps to L1 subfamily j ;

ϵ_j is the percentage of reads emanating from the L1 Subfamily j .

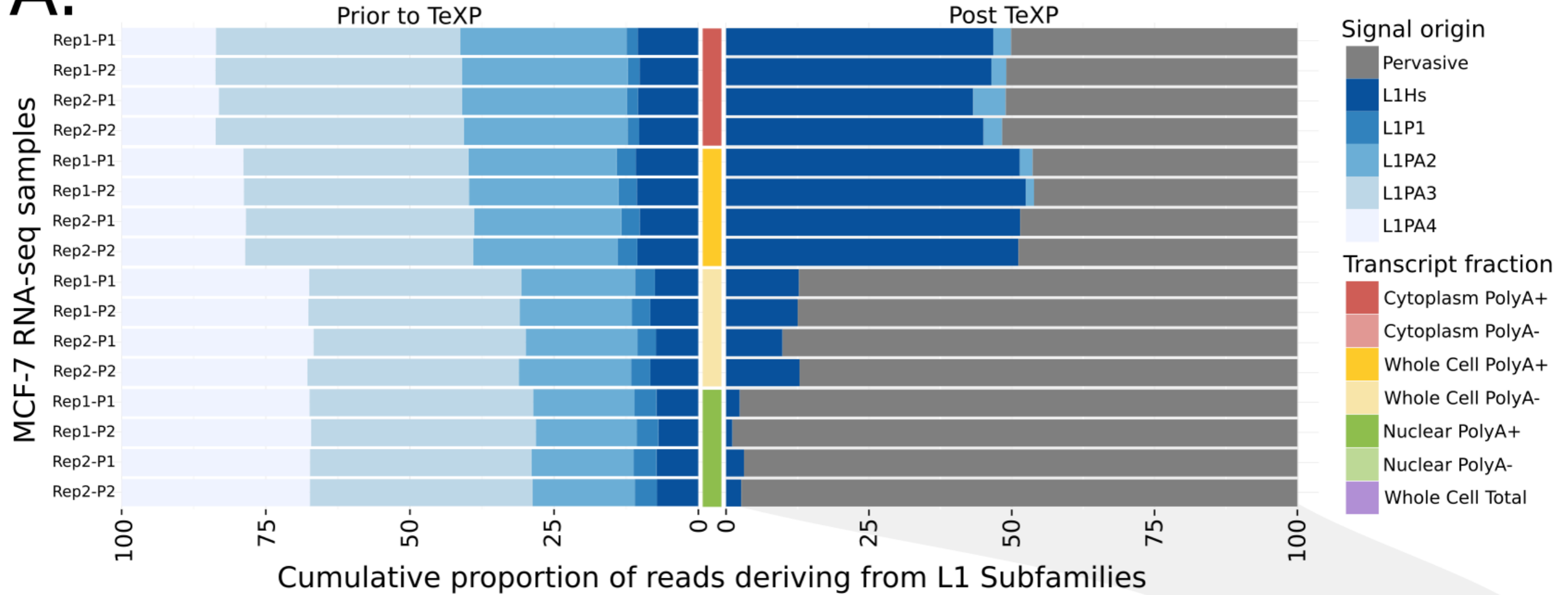


TeXP Validation

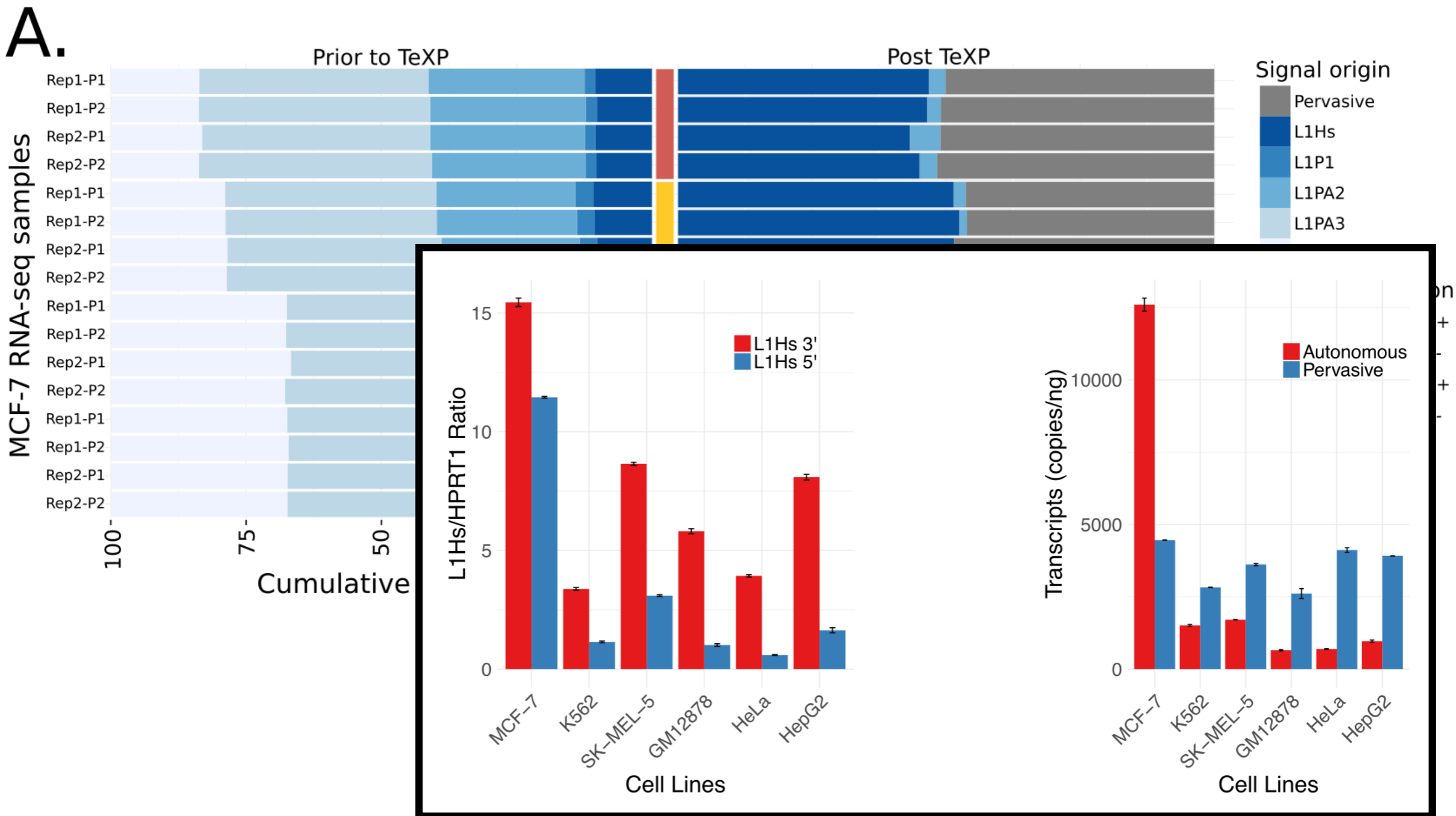


TeXP Validation

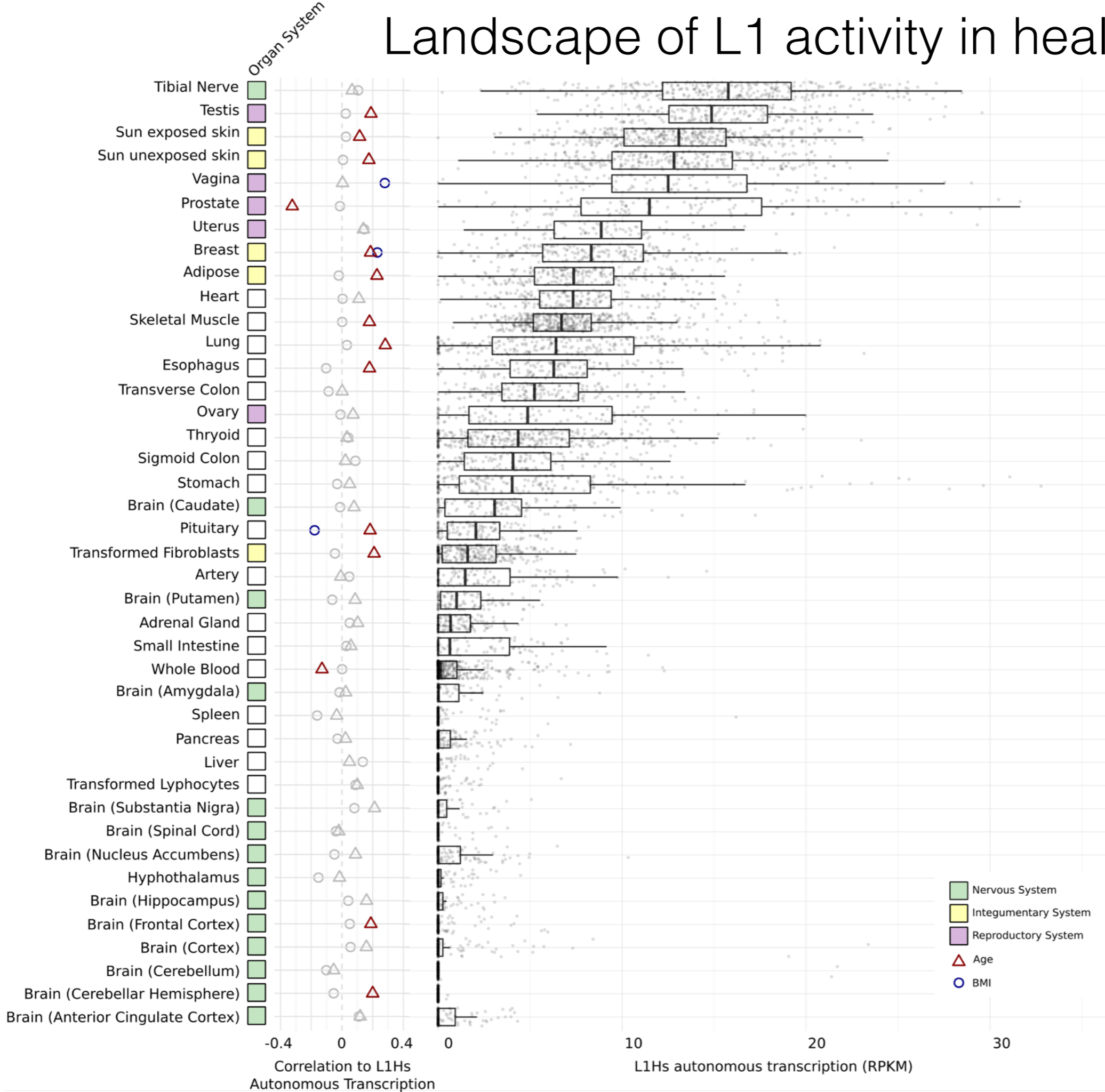
A.



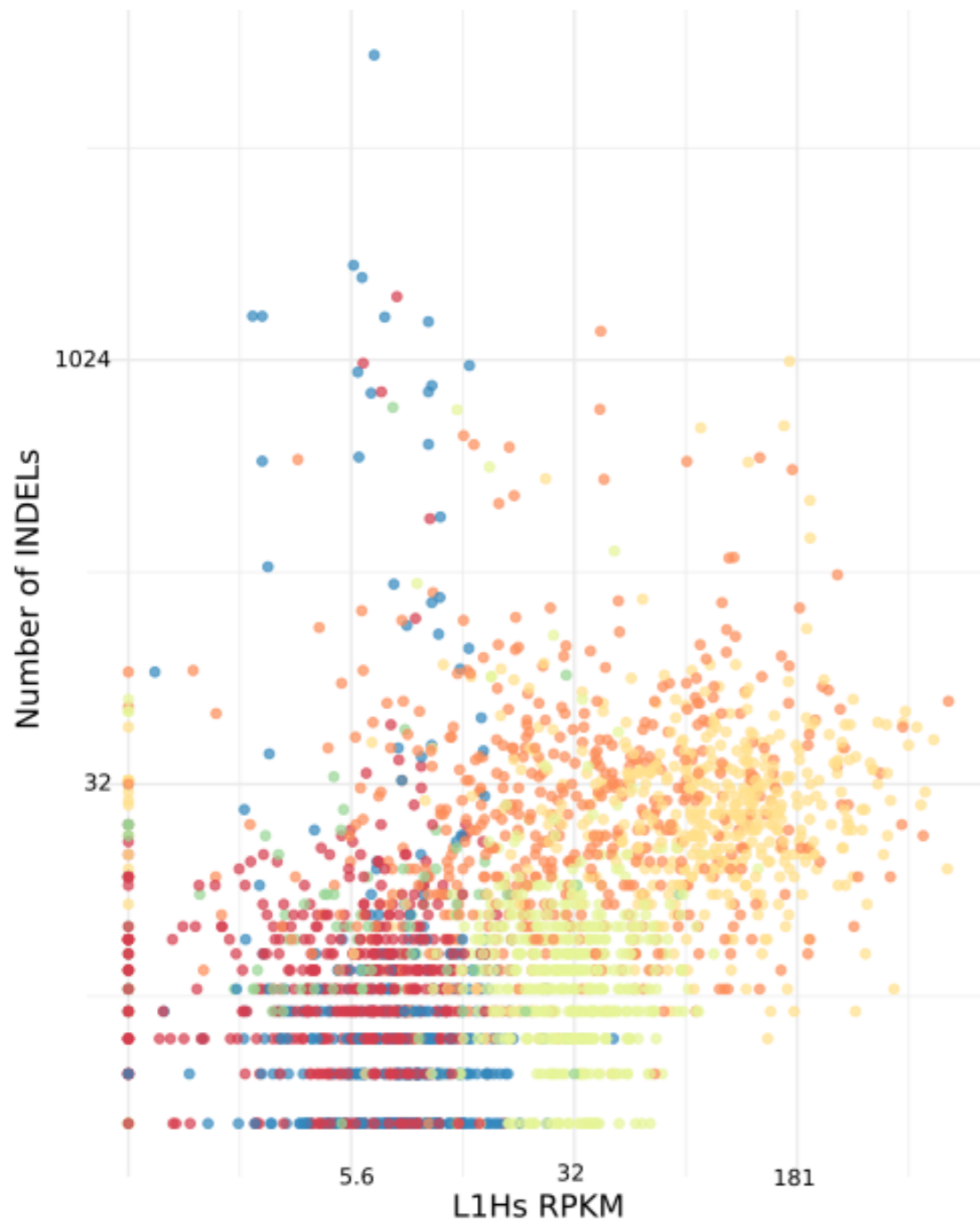
TeXP Validation



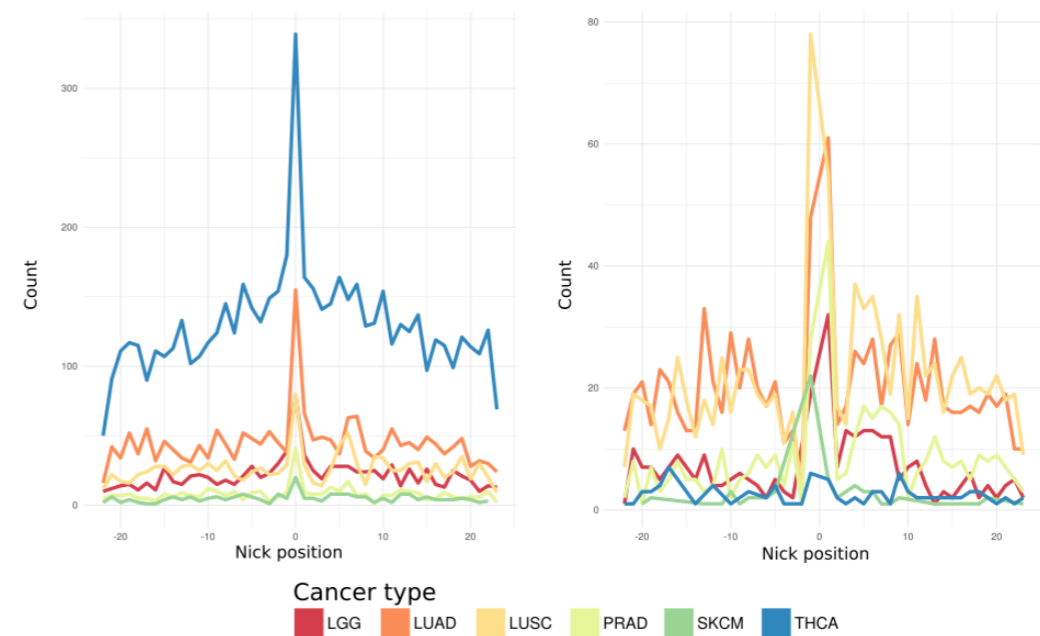
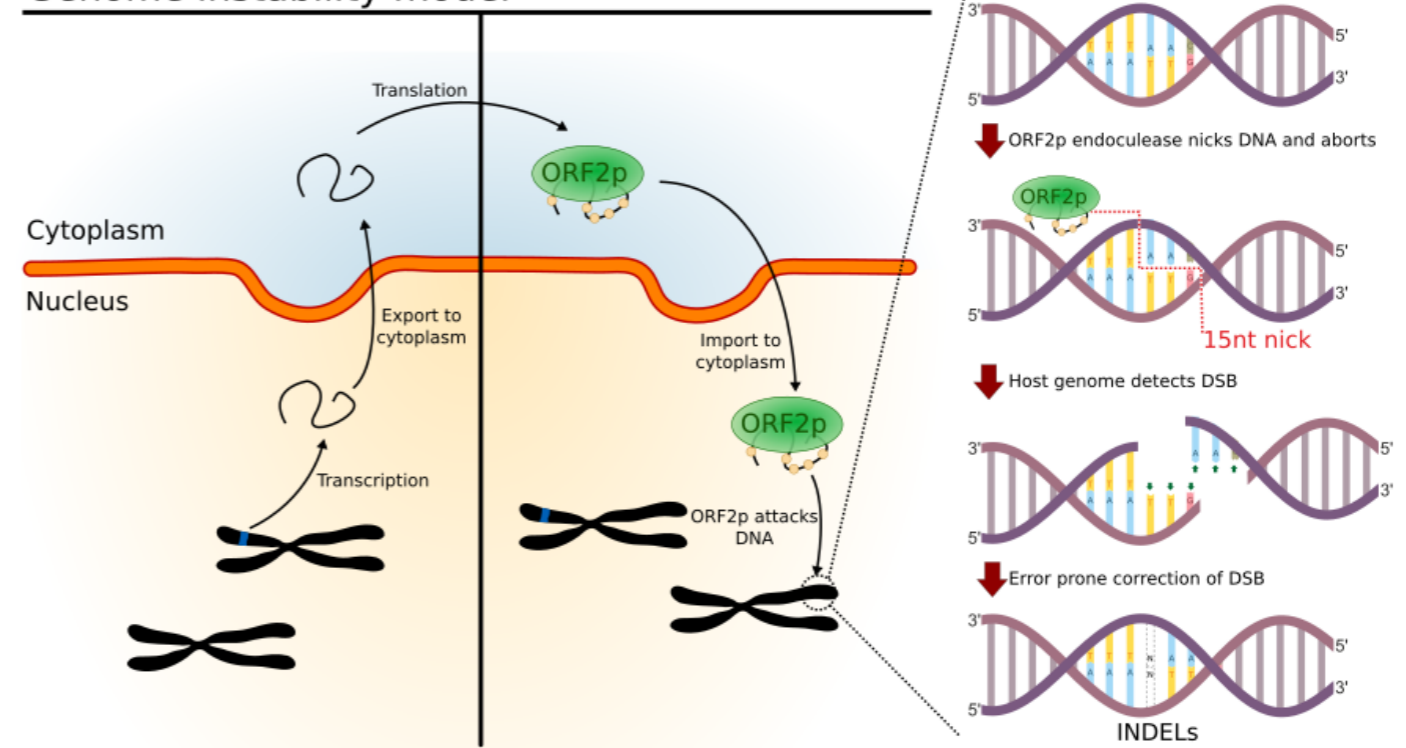
Landscape of L1 activity in healthy tissues



Correlation between L1 expression and Number of INDELS in tumors




Genome instability model



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


CANCER GENOMICS CLOUD
SEVEN BRIDGES

Log In

To access the full functionality of the CGC, log in with your eRA Commons or NIH cit account.

To access Controlled Data, you need to be approved for data access by dbGaP and must agree to use the data only in a manner consistent with your data access request. For more details click [here](#).

Having trouble? Try [resetting](#) your eRA Commons password.

LOGIN VIA ERA COMMONS

Log In without NIH authentication

Use this option if you don't have an eRA Commons or NIH cit account.

forgot?

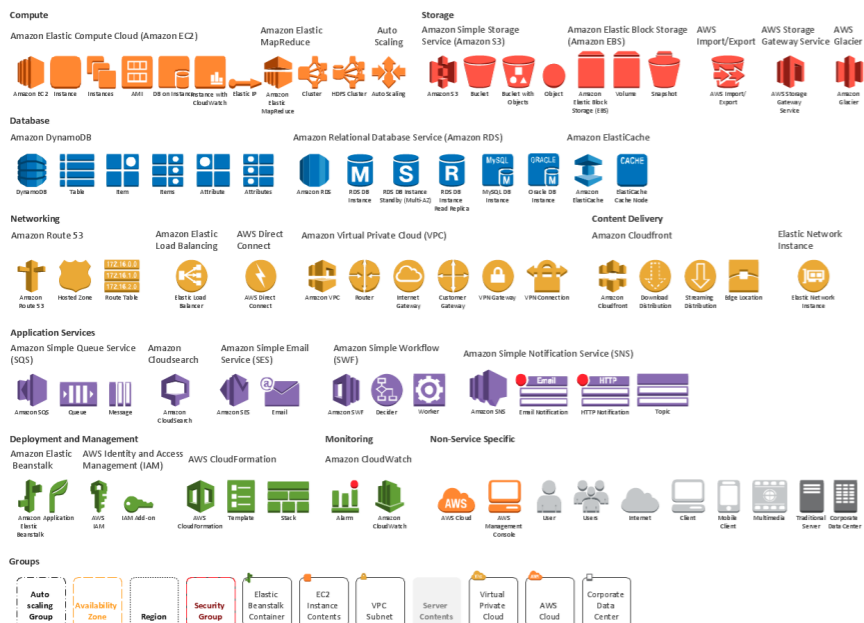
Remember me

LOGIN

New? [Create a free account](#)

Front-End

Amazon AWS



The image shows a comprehensive grid of AWS services categorized into: Compute (EC2, S3, EBS, etc.), Database (DynamoDB, RDS, ElastiCache), Networking (Route 53, VPC, CloudFront), Application Services (SQS, SES, SNS), and Deployment and Management (Elastic Beanstalk, CloudFormation, IAM). A 'Groups' section at the bottom lists various organizational structures like Auto scaling Group, Availability Zone, Region, Security Group, etc.



Amazon EC2

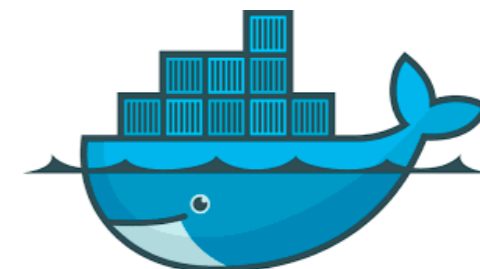


Amazon S3

Rabix

CWL

[Reproducible Analysis for Bioinformatics]



Docker

CGC Life cycle

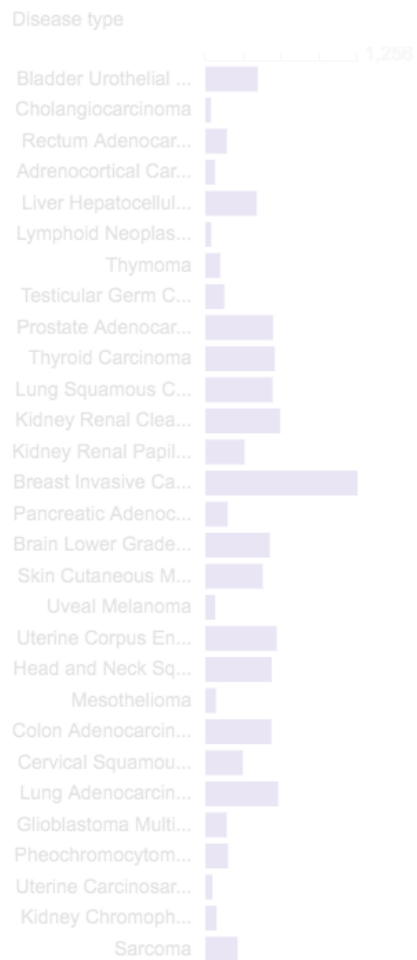


Tumor RNA-seq



Amazon S3

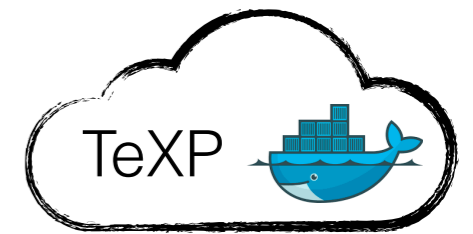
10,089



Amazon EC2



Container



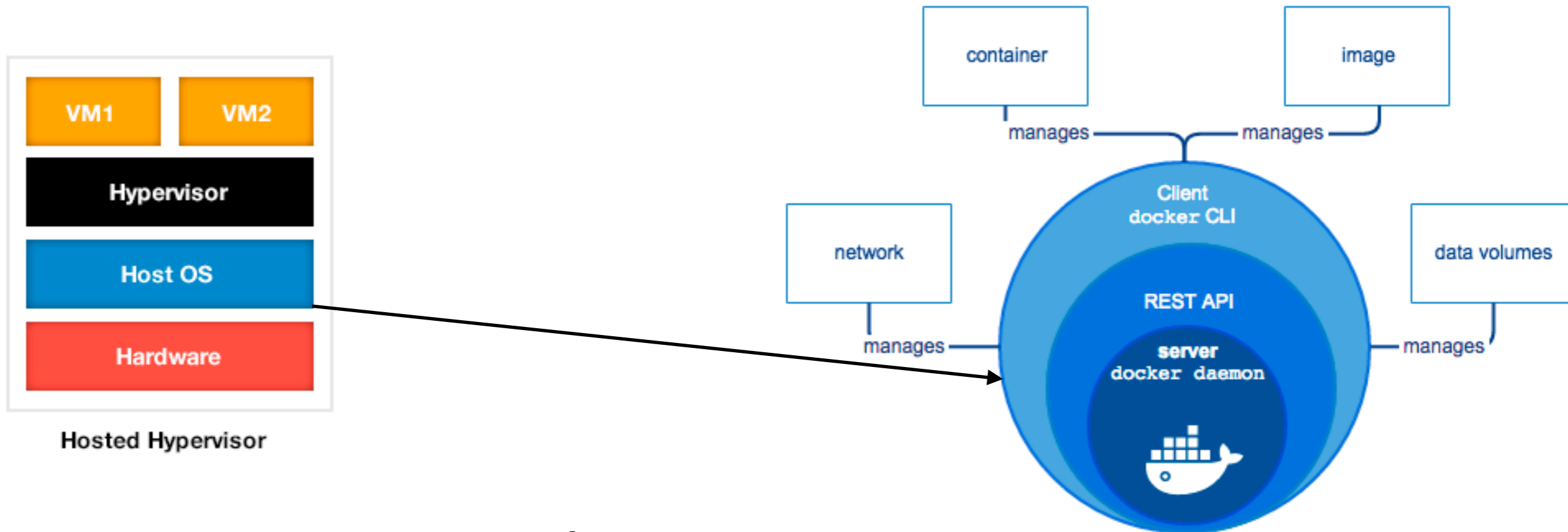
Image



Amazon S3

Result container

Docker?



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/temp: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/temp...
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>**

**cd /src ; \
 git clone <https://github.com/gersteinlab/texp.git>**

git clone <https://github.com/gersteinlab/texp.git>

WORKDIR /src/texp/

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

TeXP Dockerfile

116 commits 1 branch 1 release 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

fabiocpn Update opt.mk (compatible with docker img) Latest commit 0a5470b 3 days ago

librarian	Removed TPM quantification + Cleaning for public release	11 months ago
library	Remove plotting from lasso regression	11 months ago
Dockerfile	Whoops!	19 days ago
Makefile	Included an onliner to fix a few fastq format	19 days ago
README.md	Update README.md	5 months ago
TeXP.sh	Fixing texp path after moving to the gersteinlab repo	20 days ago
TeXP_batch.sh	Time to run it!	19 days ago
opts.mk	Update opt.mk (compatible with docker img)	3 days ago

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

```
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"]
```

docker.io + github

The screenshot shows the Docker Cloud interface for the repository `fnavarro/texp`. The top navigation bar includes the Docker Cloud logo, a plus sign, "Get Help", and the user profile "fnavarro". The breadcrumb path is "Repositories / fnavarro / texp". Below this are tabs for "General", "Tags", "Builds", "Timeline", and "Settings". A "Launch service" button is visible in the top right.

Repository Information:

- Repository: `fnavarro/texp`
- Description: TeXP is a pipeline to estimate the autonomous transcription of L1 elements based of RNA-seq data
- Last pushed: 16 days ago

Docker commands:

```
$ docker push fnavarro/texp:tagname
```

Tags:

This repository contains 5 tag(s).

Tag	Icon	Time
latest		16 days ago
1.3		17 days ago
1.2		5 months ago
1.1		a year ago
1.0		a year ago

[See all](#)

Recent builds:

- gersteinlab/texp
- Build in 'master' (0a5470bc)
- Build in 'master' (c473b17b)
- Build in 'master' (8a98f427)

[See all](#)

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

Setting up TeXP on CGC

Setting up TeXP on CGC (1)

GENERAL INPUTS OUTPUTS ADDITIONAL INFO TEST

Docker Container ⓘ

Docker Repository[:Tag]
fnavarro/teXP:latest

Resources ⓘ

CPU	Memory (MB)
1	1000

Create Files ⓘ

+ Click the plus button to add a file

Command ⓘ

Base Command

- + cwd="\$PWD";
- cd
- /src/teXP;
- ./TeXP_batch.sh
- Enter value

Stdin Stdout

Enter value Enter value

Success Codes ⓘ Temporary Fail Codes ⓘ

+ Click the plus button to add codes + Click the plus button to add codes

Arguments ⓘ

+ Click the plus button to add command line binding.



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

Setting up TeXP on CGC (2)

texp Revision 30 Try the new version of this editor. Save Run ...

GENERAL INPUTS OUTPUTS ADDITIONAL INFO TEST

Input Ports

+	0 bamfile File			
→	Prefix: -f	Separate: true	Value: Not defined	
	0 threads int			
→	Prefix: -t	Separate: true	Value: 1	
	0 output_dir string			
	Prefix: -o	Separate: true	Value: "process/"+\$job.inputs.bamfile.path.split('/')...slice(-1)[0].split('.').slice(0)[0]+"/"	
	1 output_err string			
→	Prefix: Not defined	Separate: true	Value: "2> "+\$job.inputs.bamfile.path.split('/')...slice(-1)[0].split('.').slice(0)[0]+"err"	
	1 output_log string			
	Prefix: Not defined	Separate: true	Value: "> "+\$job.inputs.bamfile.path.split('/')...slice(-1)[0].split('.').slice(0)[0]+"log"	
	2 tar string			
	Prefix: Not defined	Separate: false	Value: "; tar czvf "+\$job.inputs.bamfile.path.split('/')...slice(-1)[0]+"tar.gz process/"+\$jo...	
→	3 move string			
	Prefix: Not defined	Separate: true	Value: "mv "+\$job.inputs.bamfile.path.split('/')...slice(-1)[0]+"tar.gz "+\$job.inputs.bamfi...	

-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)

← **texp** Revision 30 ▾

Try the [new version](#) of this editor.

Save

Run

⋮

GENERAL

INPUTS

OUTPUTS

ADDITIONAL INFO

TEST



Output Ports ⓘ

+

▼ **results** *File*



Glob: `$job.inputs.bamfile.path.split('/').slice(-1)[0]+".tar.gz"`

▼ **outputerr** *File*



Glob: `$job.inputs.bamfile.path.split('/').slice(-1)[0].split('.').slice(0)[0]+".err"`

▼ **ourputlog** *File*



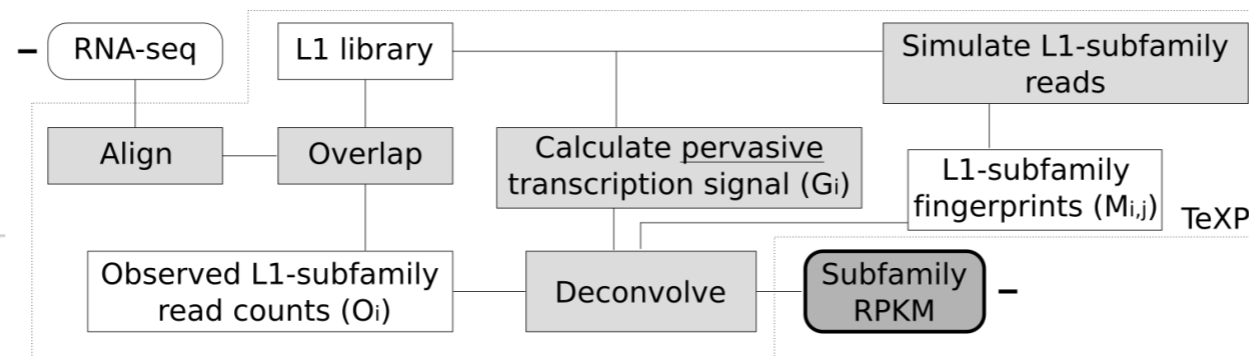
Glob: `$job.inputs.bamfile.path.split('/').slice(-1)[0].split('.').slice(0)[0]+".log"`

Setting up TeXP on CGC (4)

Create App workflow

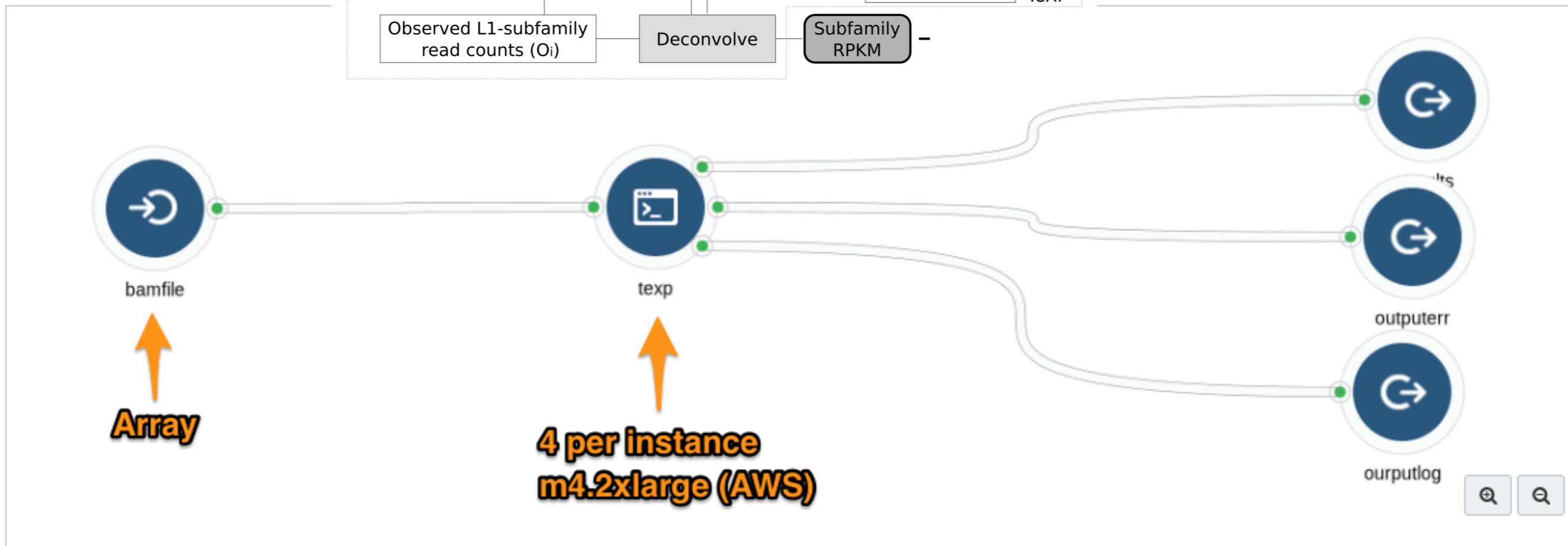
Description

No description.



```
process/UNCID_2332590/130718_UNCL2-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.err
process/UNCID_2332590/130718_UNCL2-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.log
process/UNCID_2332590/130718_UNCL2-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq.stats

process/UNCID_2332590/130718_UNCL2-SN629_0321_AC2E6WACXX_ATGTCA_L007_1.fastq:
LIHS_hg38.count
LIHS_hg38.count.corrected
LIHS_hg38.count.rpkm
LIHS_hg38.count.rpkm.corrected
LIHS_hg38.count.signal_proportions
qualityEncoding
read_length
re.bam
re.filtered.bam
re.filtered.bed
sorted.bam.bai
sorted.bam.tot
LIHS_hg38.signatures.txt
```



CGC Demo (3min)

<http://cgc.sbggenomics.com>

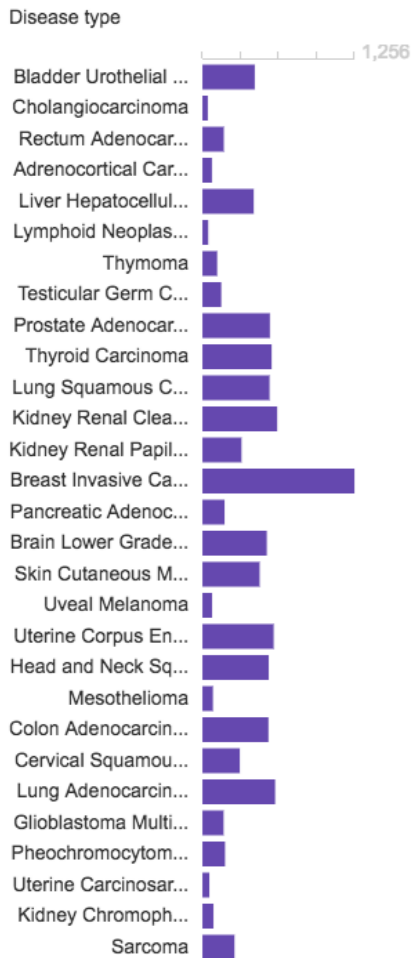
CGC Life cycle



Tumor RNA-seq



Amazon S3



10,089



Amazon EC2



TeXP Container



Image



Amazon S3

Result container

TCGA data on CGC

(if you have authorized access)

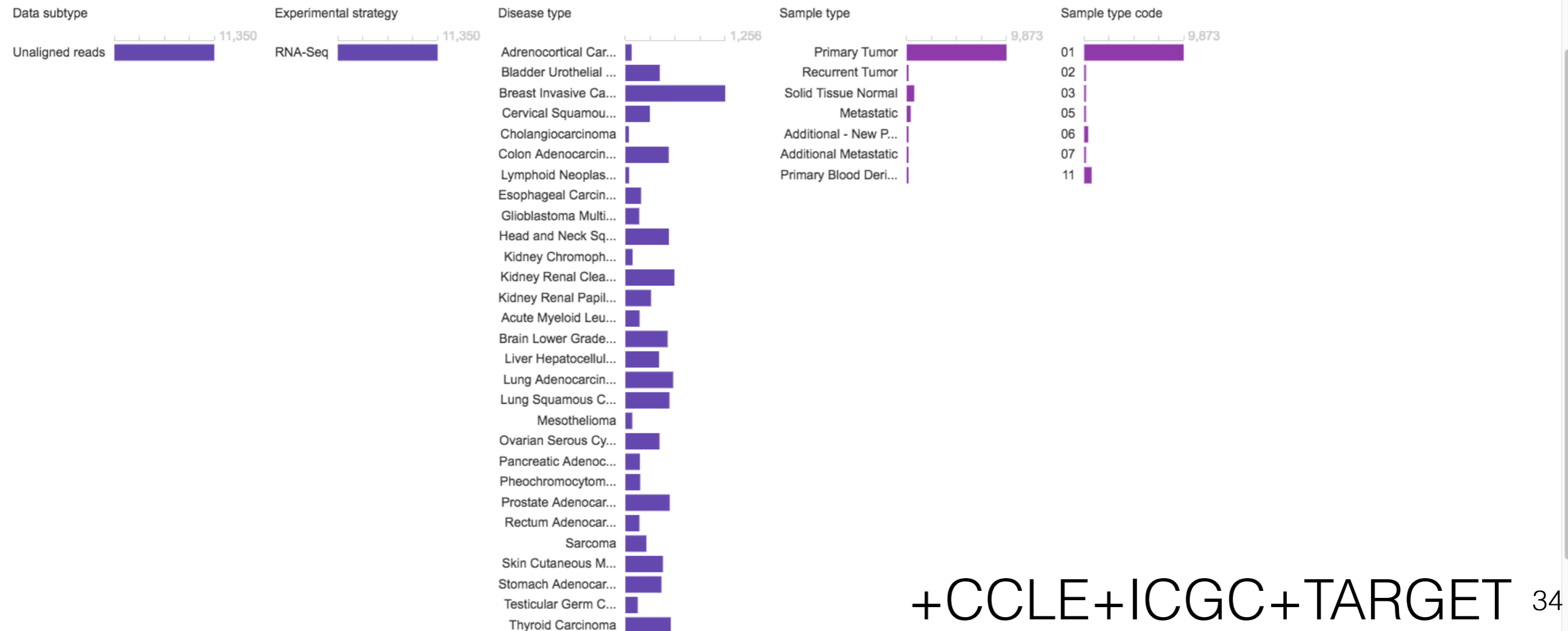
File: Data subtype (Unaligned reads), Experimental strategy (RNA-Seq), Disease type (Acute Myeloid Leukemia, Adrenocortical Carcino..., Bladder Urothelial Carc..., Brain Lower Grade Gli..., Breast Invasive Carcin..., Cervical Squamous Ce..., Cholangiocarcinoma, Colon Adenocarcinoma, Esophageal Carcinoma, Glioblastoma Multiforme, Head and Neck Squam...)

Sample: Sample type (Additional - New Primary, Additional Metastatic, Metastatic, Primary Blood Derived..., Primary Tumor, Recurrent Tumor, Solid Tissue Normal), Sample type code (01, 02, 03, 05, 06, 07, 11)

+ Add property

File: 11,350 -- Sample: 11,248 --

List Details Analytics



+CCLE+ICGC+TARGET 34

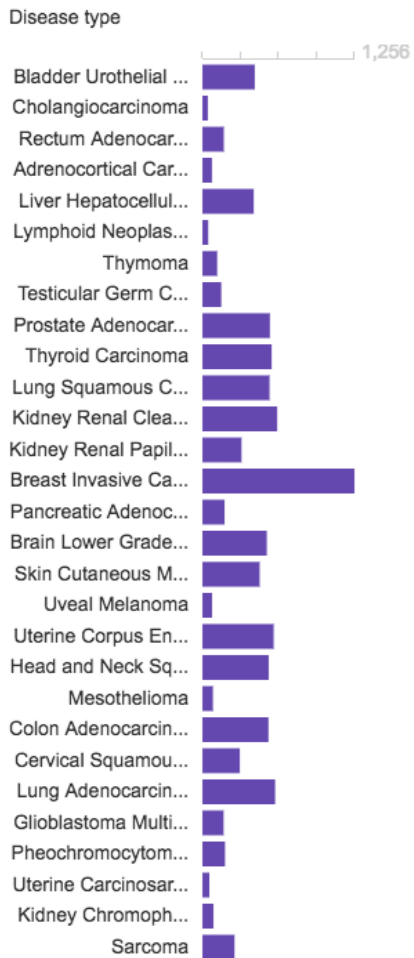
CGC Life cycle



Tumor RNA-seq



Amazon S3



10,089



Amazon EC2



Container



TeXP

Image



Amazon S3

Result container

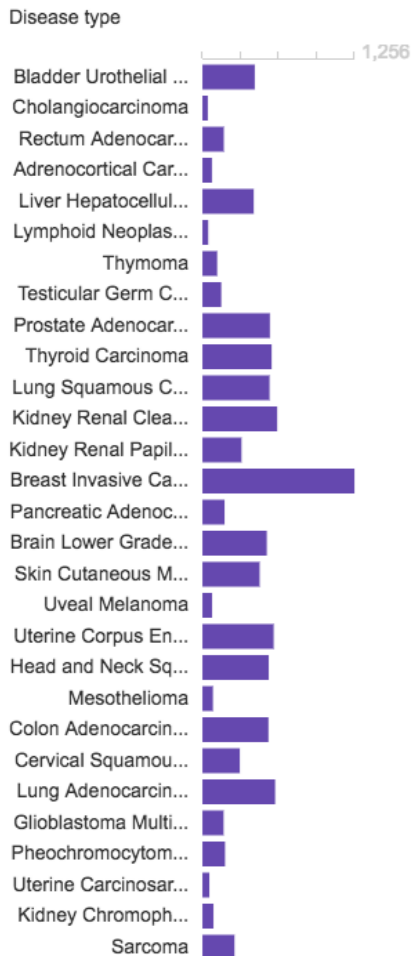
CGC Life cycle



Tumor RNA-seq



Amazon S3



10,089



Amazon EC2



Container



TeXP

Image

80 instances
= 640 proc



Amazon S3

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);

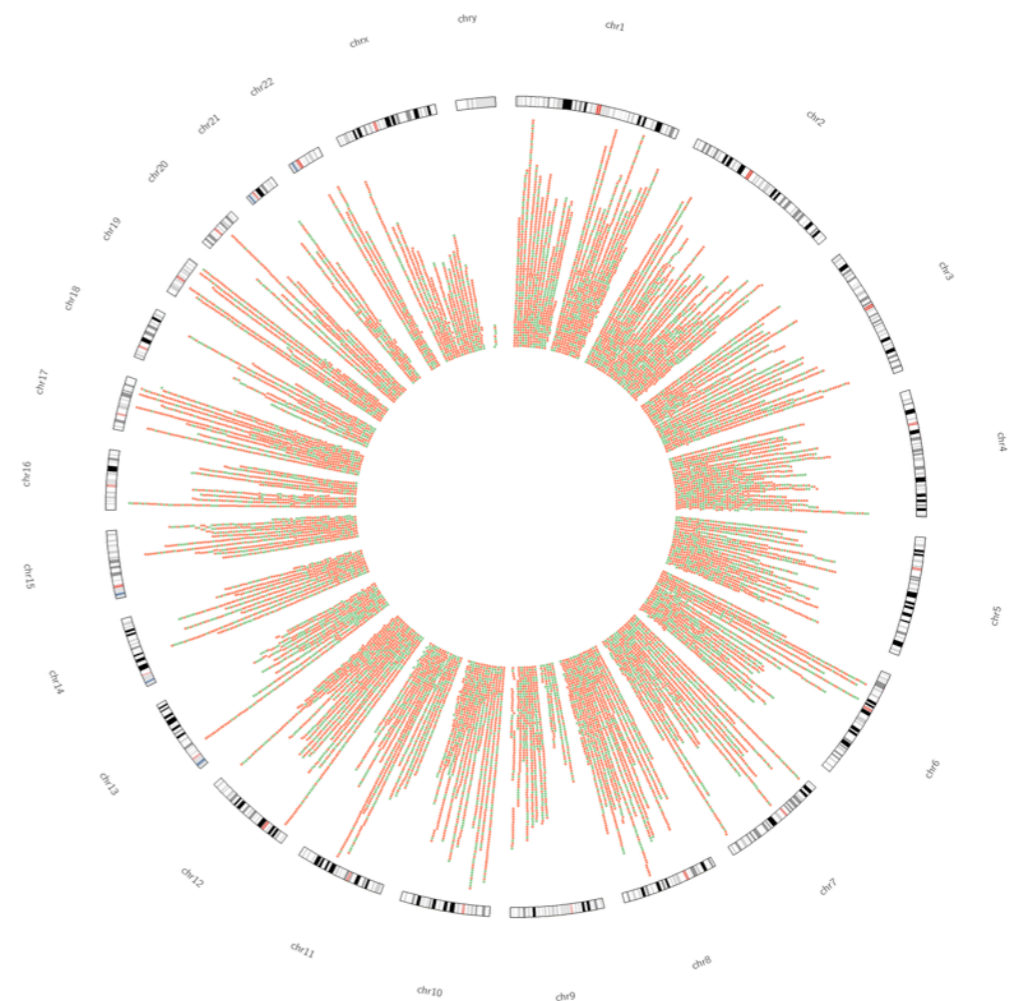
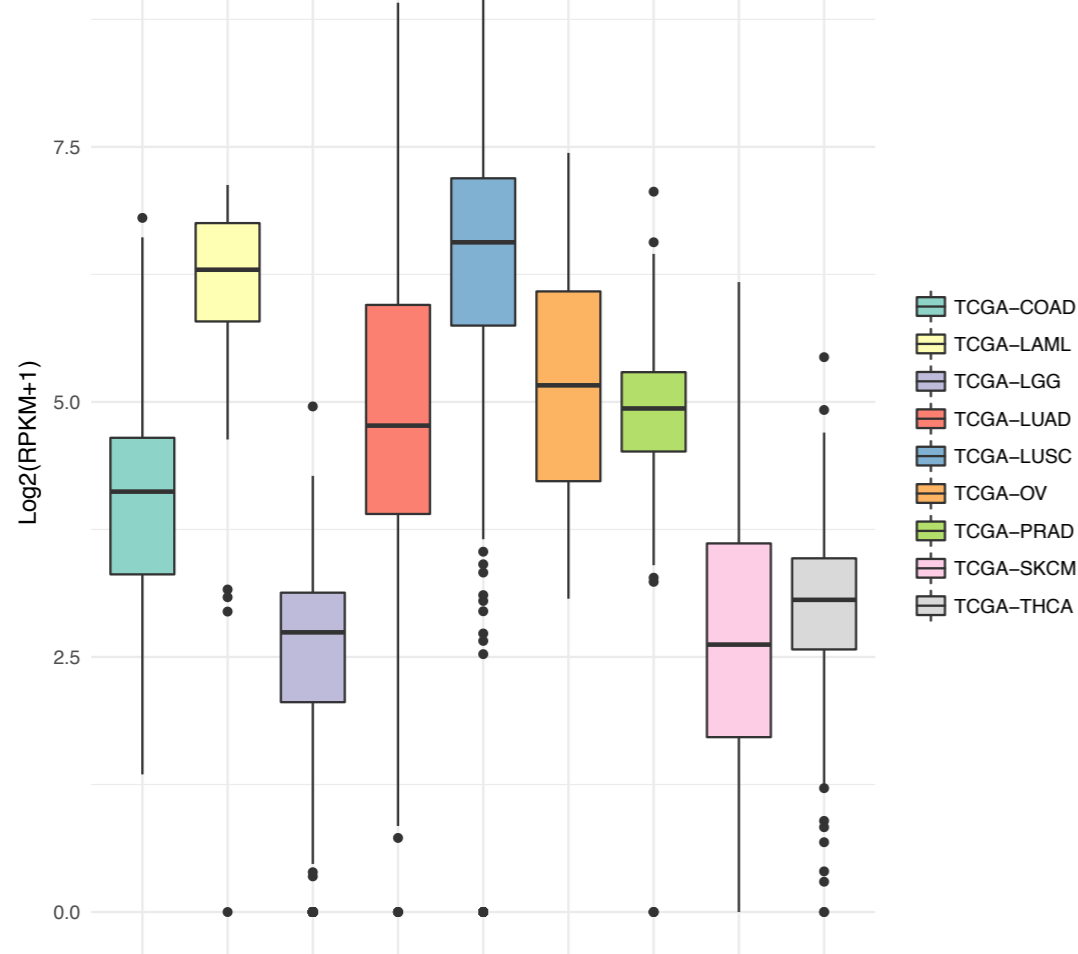
```
output_files = find_files(api, my_project, "breaks")
for i, curr_file in enumerate(output_files):
    curr_file.download(path='auto_'+curr_file.name)
    curr_file.delete()
```

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 Belfy
Eliza Cerveira
Qihui Zhu

Chengsheng Zhang

