**January 2nd, 2020**

**Non-common “admin aa account” processes note**

I have some non-common processes and I do need to document the SOP for easy access

1. Accessing the windows machine as aa-account

NIH IT support number:  866-319-4357

Go to websie: priv.nih.gov and get the password

1. Strayer University IT support: 866-610-8123 (Snowman2019!!)

**January 3rd, 2020**

**Found out an error in exomeSeq project**

I miss a variant in mouse Entpd4 gene, it turns out there are multiple regions with mm10 genome build.

BWA gives “MAQ 0” for alignment to multiple places: <https://www.biostars.org/p/83364/>

Therefore, my “bq20” criteria excludes reads with multiple locations

**To work with this, I need to retain multiple placement**

An Indian guy from Guangzhou Medical University has a post: <https://www.researchgate.net/post/How_to_derive_multiple_mapped_reads_from_a_SAM_file>

Get FLAG meeting here: <http://broadinstitute.github.io/picard/explain-flags.html>

I used BWA mem alignment, now i have reads with MAPQ 0 that i would like to extract into a new sam or bam file. Do you know if it is possible to do that with Samtools or another tool?

Untill now i tried:

samtools view -bSq 1 input.sam > output.bam

which filter out all reads with mapping quality lower then 1, but i would like to extract them to a new sam/bam file and work on them.

**January 6th, 2020**

**Working with COSMIC database**

Go to <https://www.synapse.org/#!Synapse:syn12009743> and download the signature matric I needed by selecting the file list download, adding to the project folder and downloading to the local drive.

X:\project2020\exomeSeq\cosmic-related

I happen to find out that I have a SigProfiler\_2\_5\_1\_6.zip in X:\project2019\exomSeqHCC\scripts, but it can’t be installed as an R package.

**Getting the correlation to COSMIC with cosine correlation (from Laura Riva)**

An email from Laura Riva (1/22/2019)

<https://www.linode.com/docs/tools-reference/tools/find-files-in-linux-using-the-command-line/>

**January 8th, 2020**

I keep forgetting how to find a file in linux system

Command would be “find . -name someFileName.txt”

<https://www.linode.com/docs/tools-reference/tools/find-files-in-linux-using-the-command-line/>

I also needed SignatureEstimation, which is published here: <https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#signatureestimation>

**January 9th, 2020**

**Deep learning related**

I received the loaner book “**Deep learning with Python**”, and I wanted to start the deep learning in the months to come.

To test the MNIST example on Tiguar, I need to install keras, which is simple

pip install keras –user

But, when I tried to run the code, I got error “ImportError: /lib64/libstdc++.so.6: version `GLIBCXX\_3.4.21' not found (required by /ddn/gs1/home/li11/.local/lib/python3.6/site-packages/google/protobuf/pyext/\_message.so)”.

It turns out (<https://stackoverflow.com/questions/5216399/usr-lib-libstdc-so-6-version-glibcxx-3-4-15-not-found>), I do miss it.

“strings /lib64/libstdc++.so.6 | grep GLIBC “, tells me I don’t have it!

Shall I mess up around with the installation or shall I ask Les, my colleague, for quick answer?

**Simple machine learning**

I also have the image data and would like to test the simple algorithm on it

keep forgetting how to find a file in linux system

**January 10th, 2020**

**Deep learning related**

Resume my study with Trevor and started a new thread in my workblog. I have gone way back to my old note on machine learning, it has become rusty now!

**January 17th, 2020**

For some reason, I have not document anything since January 10th. People get lazy and fall behind easily, let me see how much have I missed?

**PD-L1 related reading**

**Jupyter training from Burke Squires at NIAID**

<https://github.com/burkesquires/jupyter_training>

<https://bioinformatics.niaid.nih.gov/resources#70.1.1>

**Getting ready for MacOS**

Install Anaconda on MacOS

Jupyter training by R. Burke Squires (<https://github.com/burkesquires?tab=repositories>), he provides good training course at NIAID, and I find out that we have a local jupyter server: <https://jupyter.niehs.nih.gov/> (or course, it is ONLY available once we get behind the firewall). It provides very good portal to test any notebook for python development.

# standard setup for Python 3.5

# (switches to the environment that has also MXNet)

export PATH=/ddn/gs1/home/klimczakl/miniconda2/bin:$PATH

source activate mxnet3

# on bioinfoX

export CUDA\_VISIBLE\_DEVICES=''

# on tiguar - your preferred GPU for multiple processes

export CUDA\_VISIBLE\_DEVICES='1'

# use this to evaluate GPU memory

nvidia-smi

# use next available GPU for if your preferred GPU close to full

export CUDA\_VISIBLE\_DEVICES='2'

export CUDA\_ROOT=/ddn/gs1/home/klimczakl/miniconda2/envs/mxnet3

python

# now in Python interpreter

# alternatively include in your script

from keras import backend as K

import tensorflow as tf

config = tf.ConfigProto()

config.gpu\_options.allow\_growth=True

sess = tf.Session(config=config)

# this will give you base GPU memory allocation

nvidia-smi

The above ML environment set up will allow me to kick off “deep learning” process. He also shows me how to get the xterm windows forwarding:

+-----------------------------------------------------------------------------+

| Processes:                                                       GPU Memory |

|  GPU       PID  Type  Process name                               Usage      |

|=============================================================================|

|    1    122686    C   python                                          62MiB |

+-----------------------------------------------------------------------------+

# special setup for Python 2.7 (use based environment)

# only for legacy Keras 0.3.3, Theano 0.8.2

# requires "backend": "theano" in .keras/keras.json

export PATH=/ddn/gs1/home/klimczakl/miniconda2/bin:$PATH

export CUDA\_VISIBLE\_DEVICES='1'

export CUDA\_ROOT=/ddn/gs1/home/klimczakl/miniconda2

python -s

# now in Python interpreter

import theano

**January 21st, 2020**

Start to prepare the pathway analysis short course. Dr. Chen suggested a very good paper that shows comprehensive evaluation on existing pathway analysis mythologies: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1790-4>. Accompaning the book, the authors seem to promote their R package: <https://www.bioconductor.org/packages/release/bioc/html/ROntoTools.html>

Learning from the book:

For some reason, I have not document anything since January 10th. People get lazy and fall behind easily, let me see how much have I missed?

Reviving my code/note for the prediction model building

**January 24th, 2020**

Today is Chinese New Year’s Eve! I just got a chance to write some work note again. I will be working on the pathway analysis slides.

A good paper on evaluating the existing pathway analysis:

**January 28th, 2020**

**An R trick, one does NOT have to accept build from the source!!!**

When I try to install devtools, I am prompted with “rlang” installation, and I have to choose **“no”** to build from source options!

**January 30th, 2020**

**Apparently, I have lost some note,**

**Was able to test Nguyen’s paper but found some discrepancy**

We have been reading your paper, and found it very interesting. It is very impressive that you have done such a comprehensive evaluation on existing pathway analysis methods.

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1790-4>

When we tested the "Non-small cell lung cancer" study with 5 datasets, we found that the number of files used in your study is slightly different than those deposited. We are wondering what kind of filtering steps went into these studies. Most importantly, if you used different number of arrays in the assessment, would the results be comparable to those published in the original studies?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| GEO\_ID | Disease | Normal | Condition | Pubmed\_ID | Tissue | Platform | At\_GEO |
| GSE18842 | NSCLC | 44 | 44 | 20878980 | Lung | HG-U133 Plus 2.0 | 91 |
| GSE19188 | NSCLC | 62 | 91 | 20421987 | Lung | HG-U133 Plus 2.0 | 156 |
| GSE19804 | NSCLC | 60 | 60 | 20802022 | Lung | HG-U133 Plus 2.0 | 120 |
| GSE50627 | NSCLC | 6 | 9 | 25881239 | Lung | HuGene-10st | 15 |
| GSE6044 | NSCLC | 5 | 31 | 18992152 | Lung | Hu-Focus | 47 |

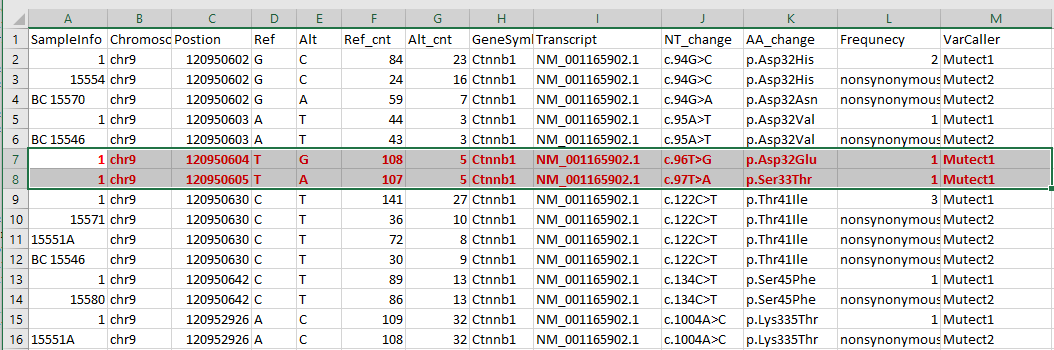
**Could potentially found the solution for the spike issues!!**

X:\project2019\exomSeqHCC\For\_Miaofei\_032019

X:\project2020\exomeSeq\shareDataWIllumina\examineVariant\historyGBEvariants\SnpEff

Between Mutect1 and Mutect2 variant calling, there is NOT much different on ctnnb1 gene. But, Mutect1 does NOT show the spikes?? It could be

1. Local realignment
2. Alignment against normal



**January 31st, 2020**

Today is the last day in January of 2020, well, it is supposed to be a good year name “2020” but ends up in a bad moment for the world. Coronavirus has caused much panic across the globe, led Dow Jones heading to a negative for the new year.

I made a wrong move for being greedy. The question is “when do I know I am greedy?” It is pretty much a binomial decision. So, just take it easy.

Well, there is much improvement on my work note as following:

**Tested a jupyter notebook I had a long time ago – successful experience.**

It has been successful so far, as I am able to test it with colab portal, I will try to run it on our internal sever today.

**Pathway evaluation paper**

I heard back from Dr. Draghici and he confirms with me about the mistakes in their paper. I have learned a lot from them though and got some data and R code forked.

**To remove a github repository I do (at least work for me)**

1. Select a repository
2. Go to “settings” and scroll down to the bottom, select “delete this repository” and follow the prompt
3. Go to local drive, and remove the folder as usual

**February 2nd, 2020**

I finish the short course on “pathway analysis” today, but I found out that there are more can be done. I am creating a thread to get comprehensive evaluation on the pathway analysis

1. I have downloaded the **GSEA** (v1.2), <https://github.com/GSEA-MSigDB/GSEA_R>, I tested the sample code, and it works!
2. Today is the last day in January of 2020, well, it is supposed to be a good year name “2020” but ends up in a bad moment for the world. Coronavirus has caused much panic across the globe, led Dow Jones heading to a negative for the new year.

**February 3rd, 2020**

I woke up early this morning and started to work. I have updated my “housekeeping” posting for record checking.

finish the short course on “pathway analysis” today, but I found out that there are more can be done. I am creating a thread to get comprehensive evaluation on the pathway analysis

**February 7th, 2020**

I create a new term:

MustARD: must (have) ambition, resolution, and determination

I'd rather be a "mustard seed", humble but not lacking the potential! In fact, most of us are like a "mustard seed", however, if we don't look down upon ourselves, sometimes, we could (hope to) make some "substantial" contribution :-)

woke up early this morning and started to work. I have updated my “housekeeping” posting for record checking.

**February 13th, 2020**

When I am working on the SEM project, I found out two bugs in the Rshinny app:

1. Initially, it was the “unique” gene count that throw things off as NOT all the vector has “unique” id, but some have the duplicate
2. It was the input data file with an extra column, but the software was looking for the count of the columns instead of “straight” fifth (5th) column.

MustARD: must (have) ambition, resolution, and determination