Tutorial on Mutational Analysis

Mutational processes, DDR mutations and Visualization

Alex Goncearenco

June 16, 2020

Goals:

1. Identify endogenic and exogenic mutational processes active in the samples using mutational data
2. Identify mutated DNA damage repair (DDR) genes that may be responsible for the endogenic mutagenesis

Outline:

1. Mutational data
2. Software and resources
3. Running the calculations
4. Visualizing the results

Data

Required (any of the following):

- GDC / TCGA somatic mutations in MAF format
- MAF or VCF files from MSKCC, ICGC or any other repository
- Or Custom-made MAF (e.g. sequenced cell lines)

Required:

- List of DDR genes

Optional:

- Annotation of signatures
- Additional sample annotations (e.g. CIMP status)

Mutation data specifics

- Germline or somatic
- “Simple” mutations vs structural changes and copy number variation
- NGS mutation callers (MuTect 1,2; Strelka, GATK, Somatic Sniper, VarScan2, VarDict, …)
- VCF files – Variant call format
- MAF files – Mutation annotation format (vcf2maf)
- Genome assembly (hg19, hg38)
- One file per sample vs multiple samples in one file

Software and resources

- MutaGene – analysis of mutational signatures and motifs
  - https://www.ncbi.nlm.nih.gov/research/mutagene/
  - https://pypi.org/project/mutagene/

- COSMIC mutational signatures
  - https://cancer.sanger.ac.uk/cosmic/signatures

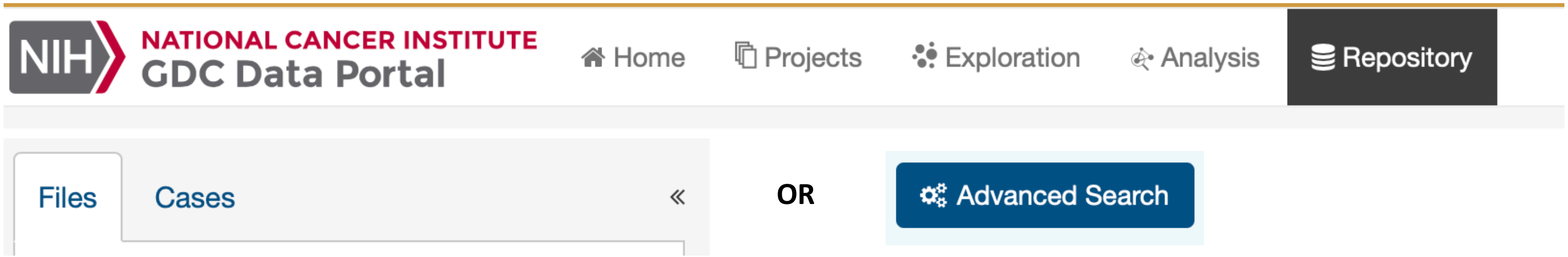
- DNA Damage Repair (DDR) genes
  - https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html
  - https://gist.github.com/neksa/df619d78cbc42c9a1abab8aa4807bfe7

- Clustergrammer
  - http://amp.pharm.mssm.edu/clustergrammer/

- Python, Jupyter notebook and Anaconda
  - https://www.anaconda.com/products/individual
  - https://github.com/MaayanLab/clustergrammer-widget

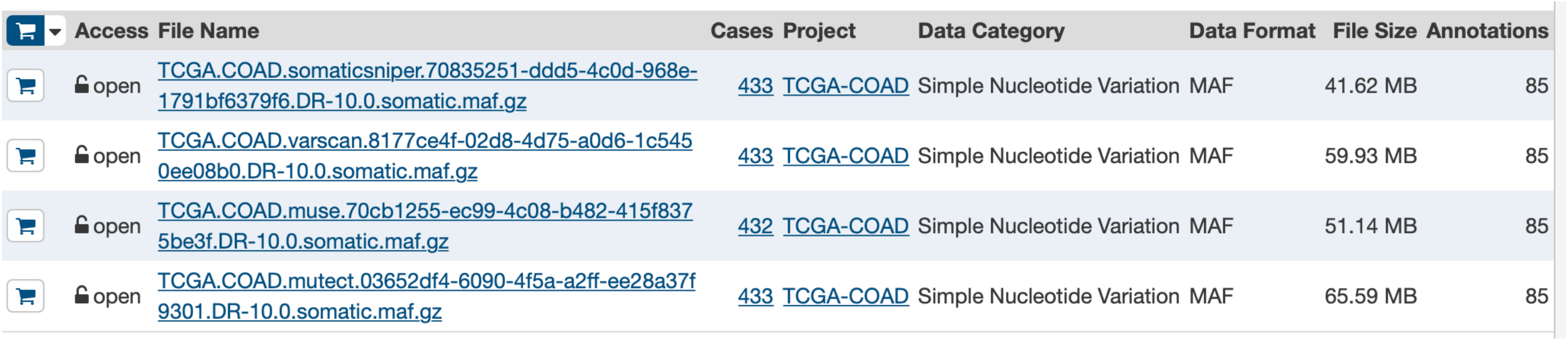
Download mutational data (TCGA)

https://portal.gdc.cancer.gov/repository



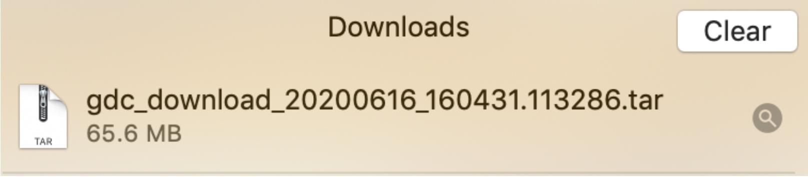
OR

```
cases.project.project_id in ["TCGA-COAD"] and files.access in ["open"] and files.data_format in ["maf"]
```

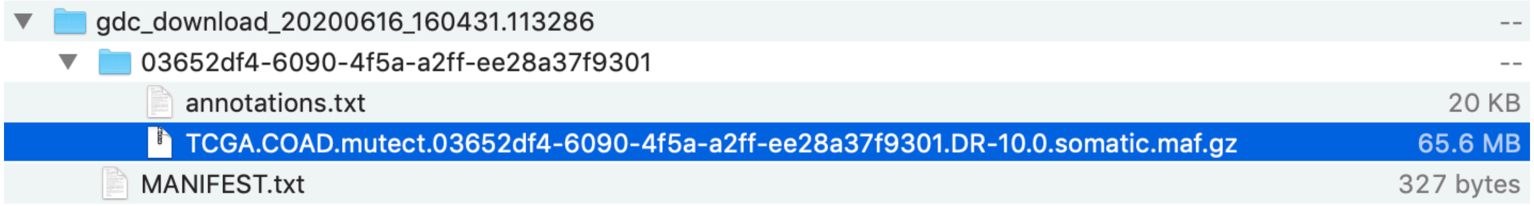


	Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
	open	TCGA.COAD.somaticsniper.70835251-ddd5-4c0d-968e-1791bf6379f6.DR-10.0.somatic.maf.gz	433	TCGA-COAD	Simple Nucleotide Variation	MAF	41.62 MB	85
	open	TCGA.COAD.varscan.8177ce4f-02d8-4d75-a0d6-1c5450ee08b0.DR-10.0.somatic.maf.gz	433	TCGA-COAD	Simple Nucleotide Variation	MAF	59.93 MB	85
	open	TCGA.COAD.muse.70cb1255-ec99-4c08-b482-415f8375be3f.DR-10.0.somatic.maf.gz	432	TCGA-COAD	Simple Nucleotide Variation	MAF	51.14 MB	85
	open	TCGA.COAD.mutect.03652df4-6090-4f5a-a2ff-ee28a37f9301.DR-10.0.somatic.maf.gz	433	TCGA-COAD	Simple Nucleotide Variation	MAF	65.59 MB	85

Download and unpack files



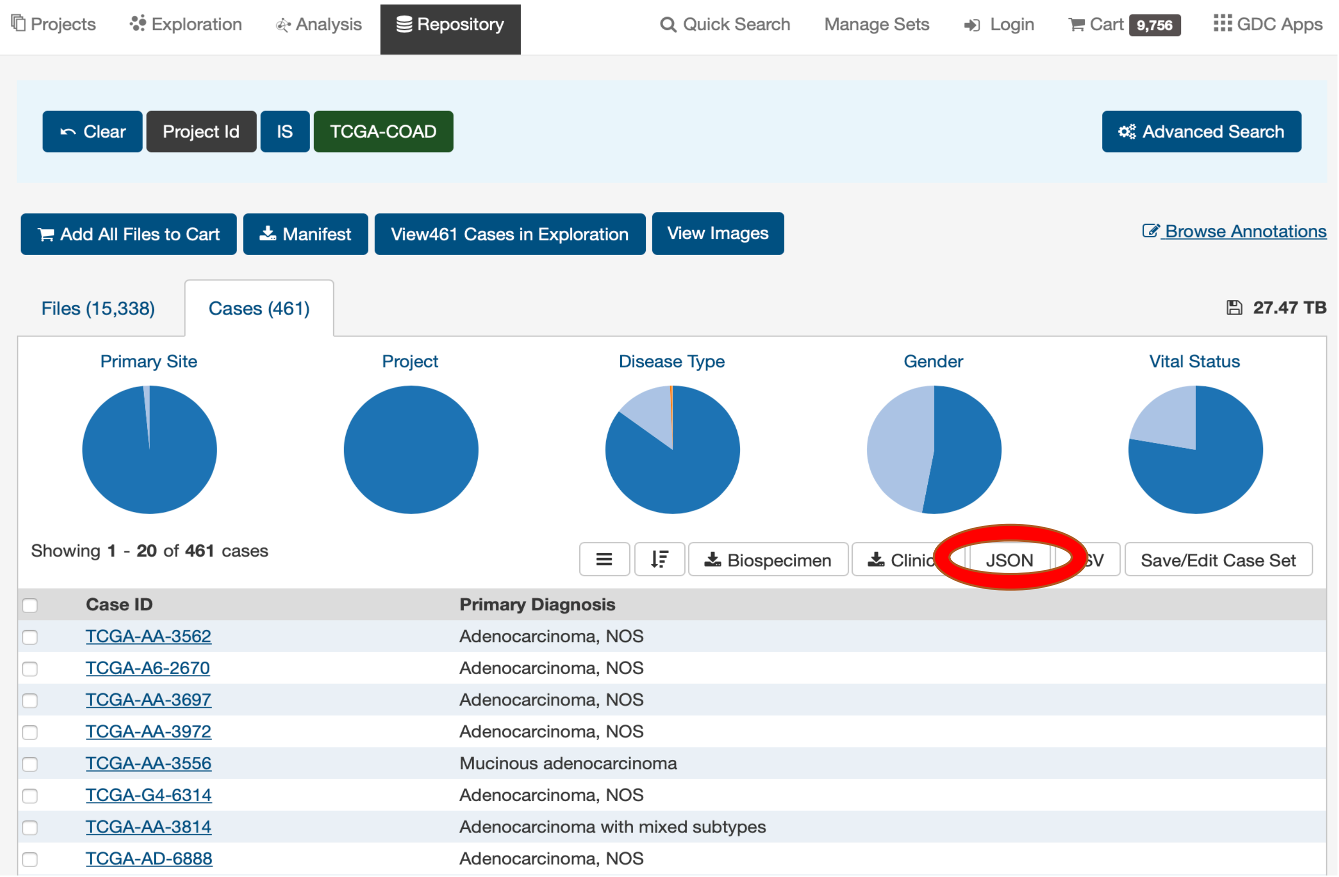
gdc_download_20200616_160431.113286.tar
65.6 MB



gdc_download_20200616_160431.113286	--
03652df4-6090-4f5a-a2ff-ee28a37f9301	--
annotations.txt	20 KB
TCGA.COAD.mutect.03652df4-6090-4f5a-a2ff-ee28a37f9301.DR-10.0.somatic.maf.gz	65.6 MB
MANIFEST.txt	327 bytes

```
gunzip TCGA.COAD.mutect.03652df4-6090-4f5a-a2ff-ee28a37f9301.DR-10.0.somatic.maf.gz
```

Download metadata (TCGA)

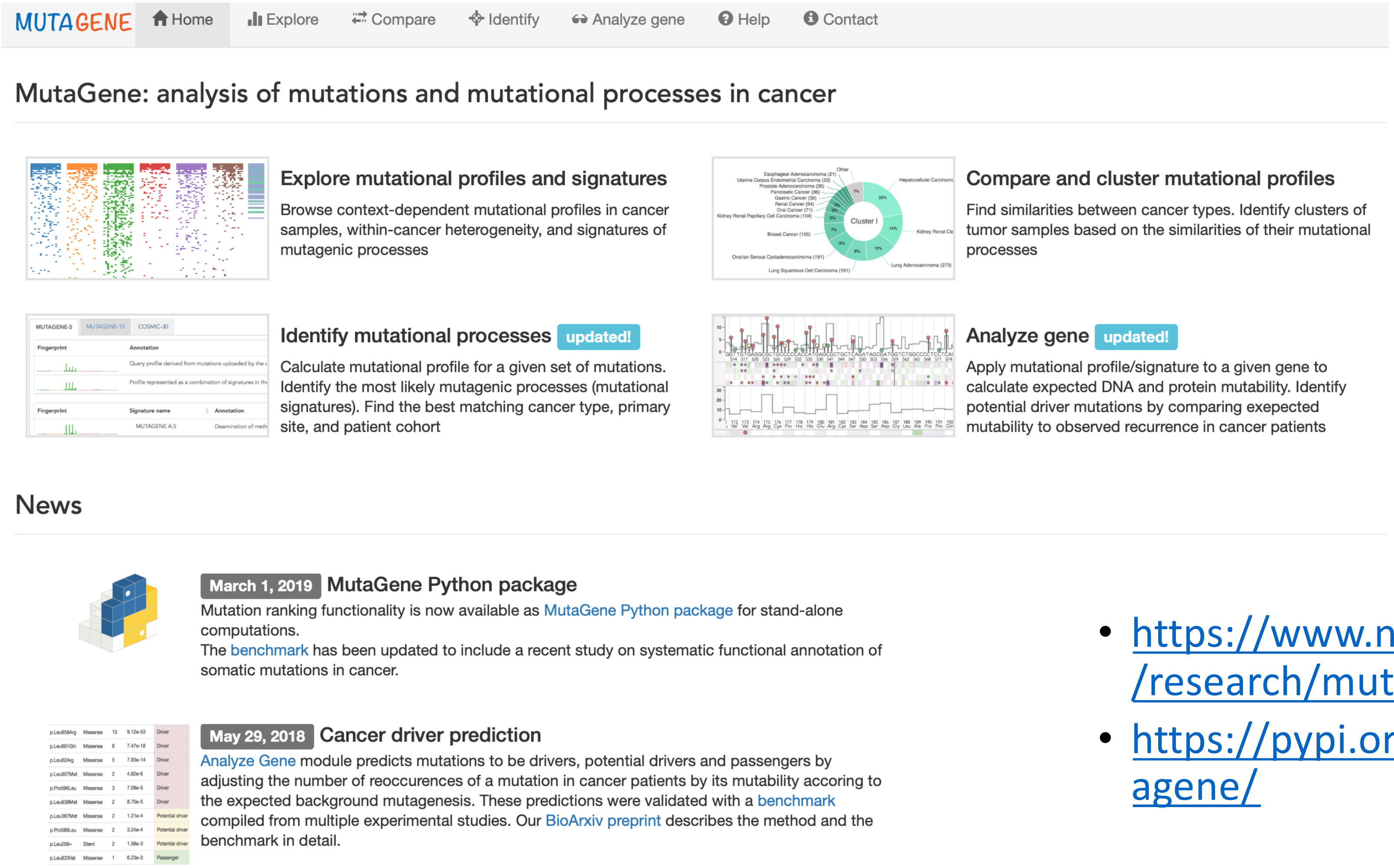


Showing **1** - **20** of **461** cases

Case ID	Primary Diagnosis
TCGA-AA-3562	Adenocarcinoma, NOS
TCGA-A6-2670	Adenocarcinoma, NOS
TCGA-AA-3697	Adenocarcinoma, NOS
TCGA-AA-3972	Adenocarcinoma, NOS
TCGA-AA-3556	Mucinous adenocarcinoma
TCGA-G4-6314	Adenocarcinoma, NOS
TCGA-AA-3814	Adenocarcinoma with mixed subtypes
TCGA-AD-6888	Adenocarcinoma, NOS

Download JSON and save as TCGA_COAD_cases.json

What is MutaGene?



MutaGene: analysis of mutations and mutational processes in cancer

Explore mutational profiles and signatures

Browse context-dependent mutational profiles in cancer samples, within-cancer heterogeneity, and signatures of mutagenic processes

Compare and cluster mutational profiles

Find similarities between cancer types. Identify clusters of tumor samples based on the similarities of their mutational processes

Identify mutational processes updated!

Calculate mutational profile for a given set of mutations. Identify the most likely mutagenic processes (mutational signatures). Find the best matching cancer type, primary site, and patient cohort

Analyze gene updated!

Apply mutational profile/signature to a given gene to calculate expected DNA and protein mutability. Identify potential driver mutations by comparing exepected mutability to observed recurrence in cancer patients

News

March 1, 2019 MutaGene Python package

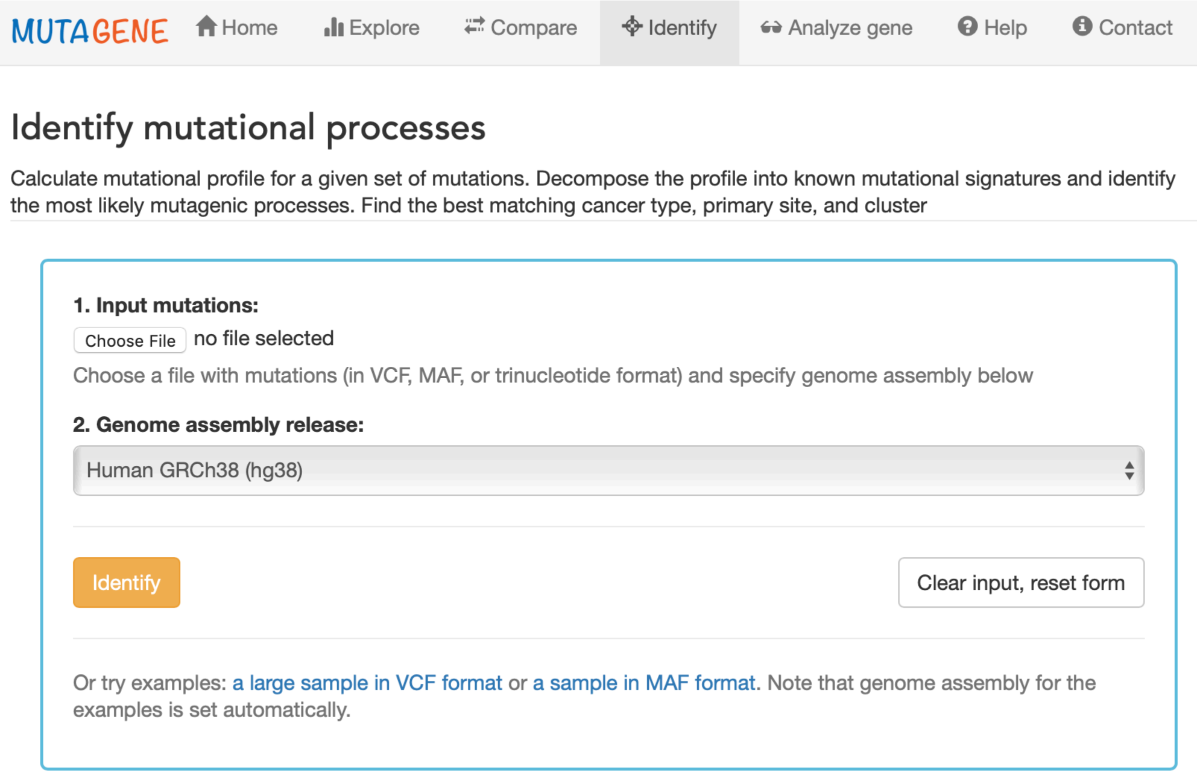
Mutation ranking functionality is now available as MutaGene Python package for stand-alone computations.
The benchmark has been updated to include a recent study on systematic functional annotation of somatic mutations in cancer.

May 29, 2018 Cancer driver prediction

Analyze Gene module predicts mutations to be drivers, potential drivers and passengers by adjusting the number of reoccurences of a mutation in cancer patients by its mutability accoring to the expected background mutagenesis. These predictions were validated with a benchmark compiled from multiple experimental studies. Our BioArxiv preprint describes the method and the benchmark in detail.

- https://www.ncbi.nlm.nih.gov/research/mutagene/
- https://pypi.org/project/mutagene/

Identify mutational processes in mutagene



Identify mutational processes

Calculate mutational profile for a given set of mutations. Decompose the profile into known mutational signatures and identify the most likely mutagenic processes. Find the best matching cancer type, primary site, and cluster

1. Input mutations:

Choose File no file selected

Choose a file with mutations (in VCF, MAF, or trinucleotide format) and specify genome assembly below

2. Genome assembly release:

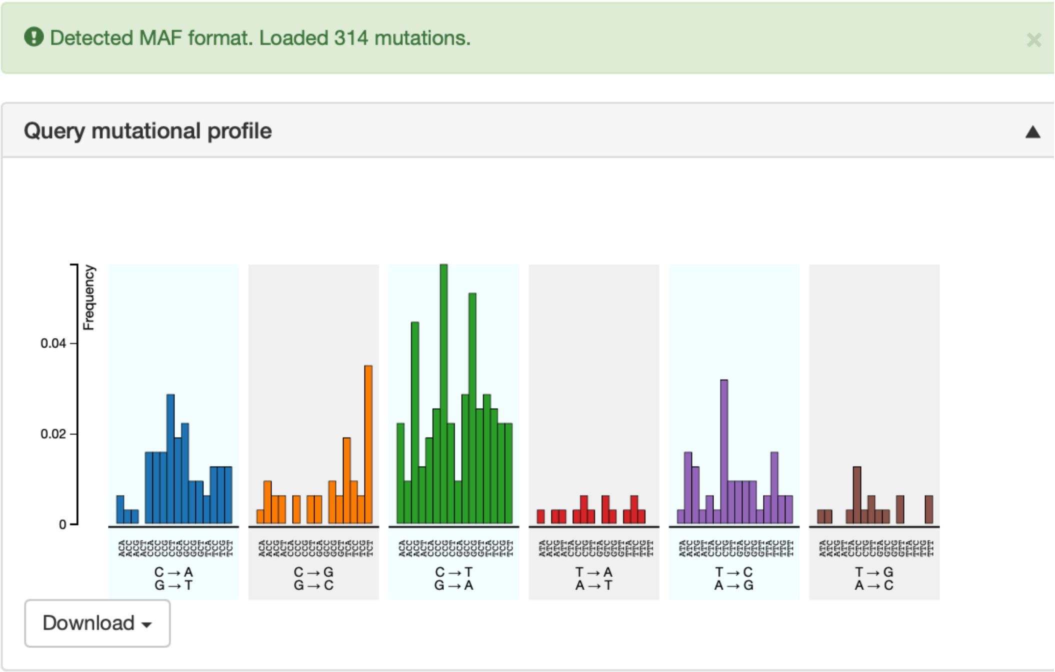
Human GRCh38 (hg38)

Identify

Clear input, reset form

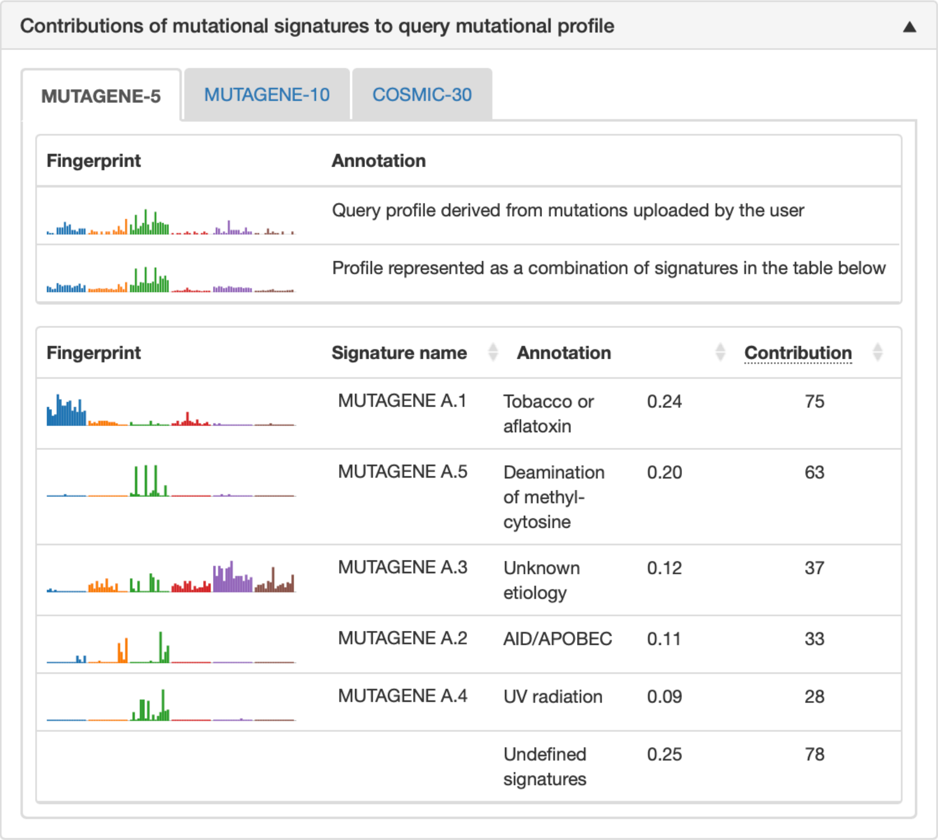
Or try examples: a large sample in VCF format or a sample in MAF format. Note that genome assembly for the examples is set automatically.

Identify mutational processes in mutagene



Detected MAF format. Loaded 314 mutations.

Query mutational profile



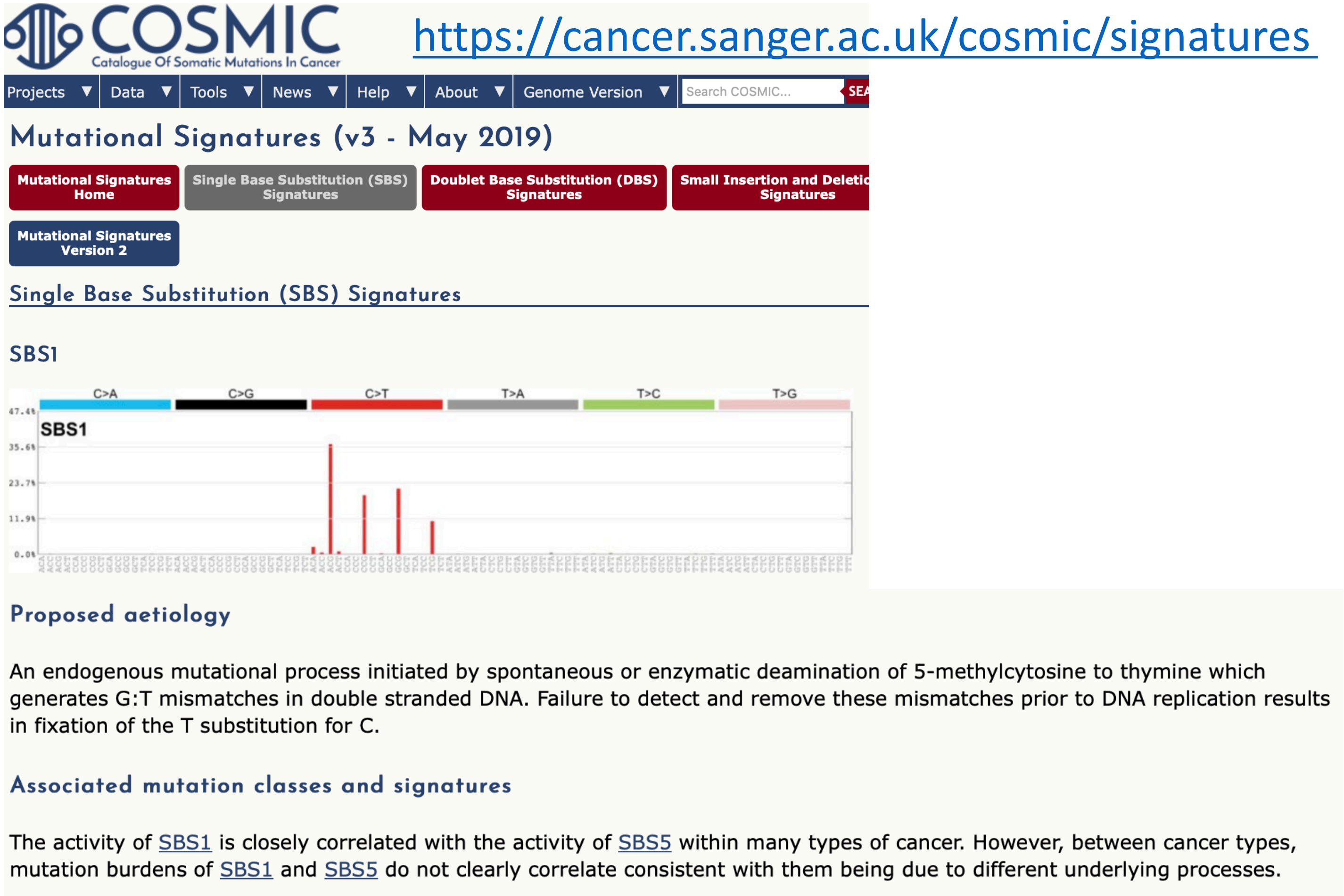
Contributions of mutational signatures to query mutational profile

Fingerprint	Annotation
	Query profile derived from mutations uploaded by the user
	Profile represented as a combination of signatures in the table below

Fingerprint	Signature name	Annotation		Contribution
	MUTAGENE A.1	Tobacco or aflatoxin	0.24	75
	MUTAGENE A.5	Deamination of methyl-cytosine	0.20	63
	MUTAGENE A.3	Unknown etiology	0.12	37
	MUTAGENE A.2	AID/APOBEC	0.11	33
	MUTAGENE A.4	UV radiation	0.09	28
		Undefined signatures	0.25	78

What are COSMIC signatures?

https://cancer.sanger.ac.uk/cosmic/signatures


COSMIC
Catalogue Of Somatic Mutations In Cancer
Projects ▼
Data ▼
Tools ▼
News ▼
Help ▼
About ▼
Genome Version ▼
Search COSMIC...
SEA
Mutational Signatures Home
Single Base Substitution (SBS) Signatures
Doublet Base Substitution (DBS) Signatures
Small Insertion and Deletio Signatures
Mutational Signatures Version 2
C>A
C>G
C>T
T>A
T>C
T>G
SBS1
47.4%
35.6%
23.7%
11.9%
0.0%

Mutational Signatures (v3 - May 2019)

Single Base Substitution (SBS) Signatures

SBS1

Proposed aetiology

An endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates G:T mismatches in double stranded DNA. Failure to detect and remove these mismatches prior to DNA replication results in fixation of the T substitution for C.

Associated mutation classes and signatures

The activity of SBS1 is closely correlated with the activity of SBS5 within many types of cancer. However, between cancer types, mutation burdens of SBS1 and SBS5 do not clearly correlate consistent with them being due to different underlying processes.

What are COSMIC signatures?

- '1': 5mC>T
- '2': AID/APOBEC
- '13': AID/APOBEC
- '3': HR DDR defect
- '4': Tobacco
- '10a': Polymerase epsilon
- '10b': Polymerase epsilon
- '14': Polymerase epsilon + MMR
- '15': MMR
- '7b': UV light
- '18': reactive oxygen species
- '24': guanine repair by NER
- '5': Unknown
- '6': DDR MSI
- '19': Unknown
- '20': MMR, POLD1
- '21': MMR
- '26': MMR
- '32': Azathioprine
- '44': MMR
- '46': Sequencing artifact
- '54': Sequencing artifact
- '84': AID

Download COSMICv3_labels.yml:
https://gist.github.com/neksa/089af7bea566f76ceabd7337b497bb96

What are DDR genes?

- DNA Damage Repair (DDR) genes
    - https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html

Download and save DDR_genes.yml file:
https://gist.github.com/neksa/df619d78cbc42c9a1abab8aa4807bfe7

MMR:
- MSH2
- MSH3
- MSH6
- MLH1
- PMS2
- MSH4
- MSH5
- MLH3
- PMS1
- PMS2P3
- PMS2L3

PARP:
- PARP1
- ADPRT
- PARP2
- ADPRTL2
- PARP3
- ADPRTL3

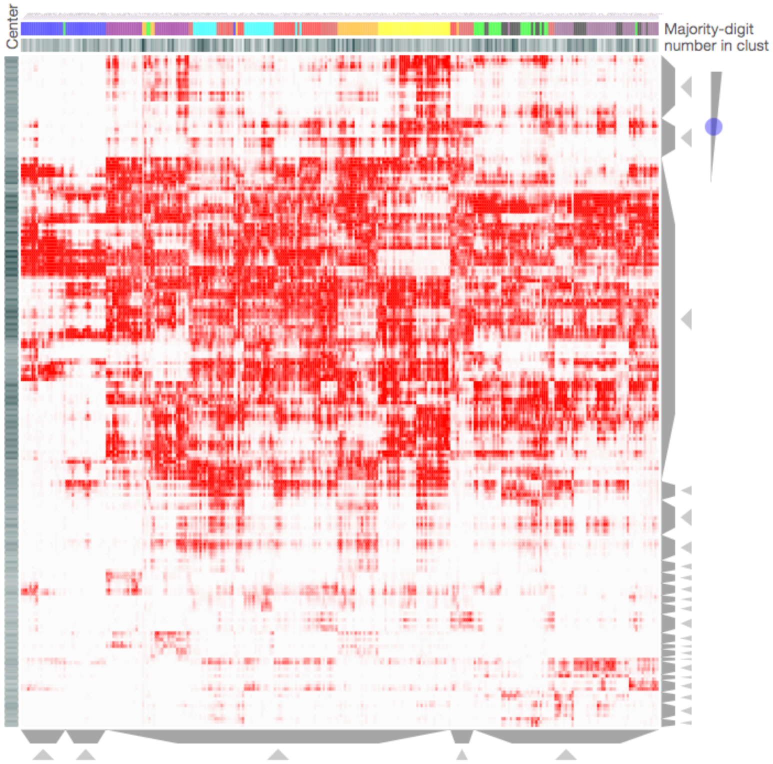
Direct-reversal:
- MGMT
- ALKBH2
- ABH2
- ALKBH3
- DEPC1

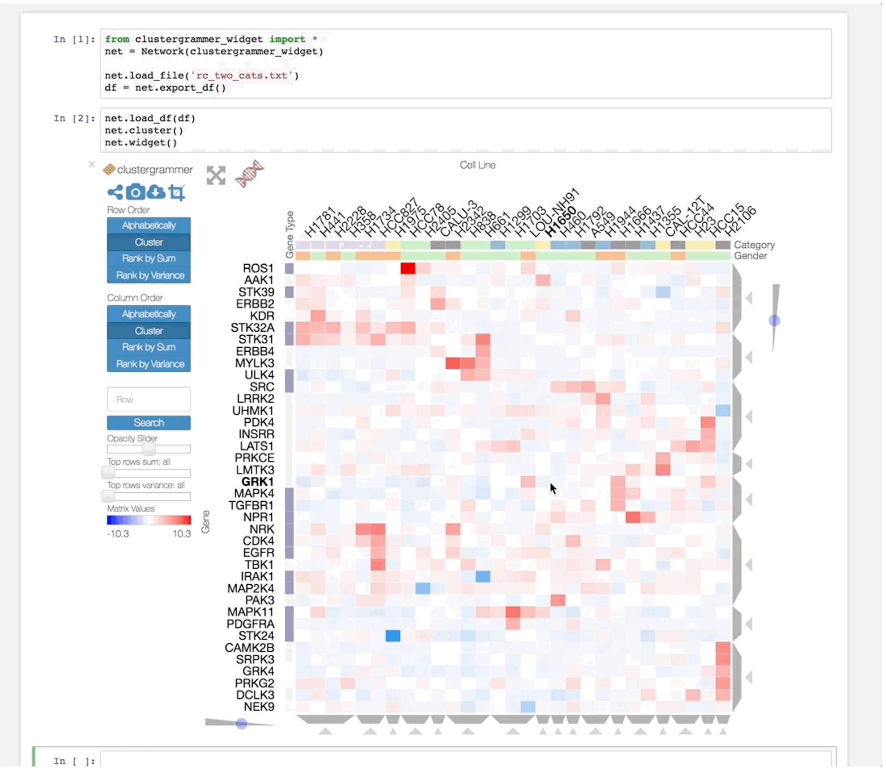
Topo-crosslinks:
- TDP1
- TDP2
- TTRAP

Consider Missense or Nonsense mutations

Do not differentiate between mutations

What is Clustergrammer?





```
In [1]: from clustergrammer_widget import *
        net = Network(clustergrammer_widget)

        net.load_file('rc_two_cats.txt')
        df = net.export_df()

In [2]: net.load_df(df)
        net.cluster()
        net.widget()
```

clustergrammer
Cell Line
Row Order
Alphabetically
Cluster
Rank by Sum
Rank by Variance
Column Order
Alphabetically
Cluster
Rank by Sum
Rank by Variance
Row
Search
Opacity Slider
Top rows sum: all
Top rows variance: all
Matrix Values
-10.3
10.3
Gene Type
Gene
Category
Gender
H1781 H441 H2228 H358 H1734 HCC827 H1975 HCC78 H2405 CALU-3 H2342 H838 H661 H1299 H1703 LOU-NH91 H1650 H460 H1792 A549 H1944 H1666 H1437 H1355 CAL-12T HCC44 H23 HCC15 H2106
ROS1 AAK1 STK39 ERBB2 KDR STK32A STK31 ERBB4 MYLK3 ULK4 SRC LRRK2 UHMK1 PDK4 INSRR LATS1 PRKCE LMTK3 GRK1 MAPK4 TGFBR1 NPR1 NRK CDK4 EGFR TBK1 IRAK1 MAP2K4 PAK3 MAPK11 PDGFRA STK24 CAMK2B SRPK3 GRK4 PRKG2 DCLK3 NEK9

```
In [ ]:
```

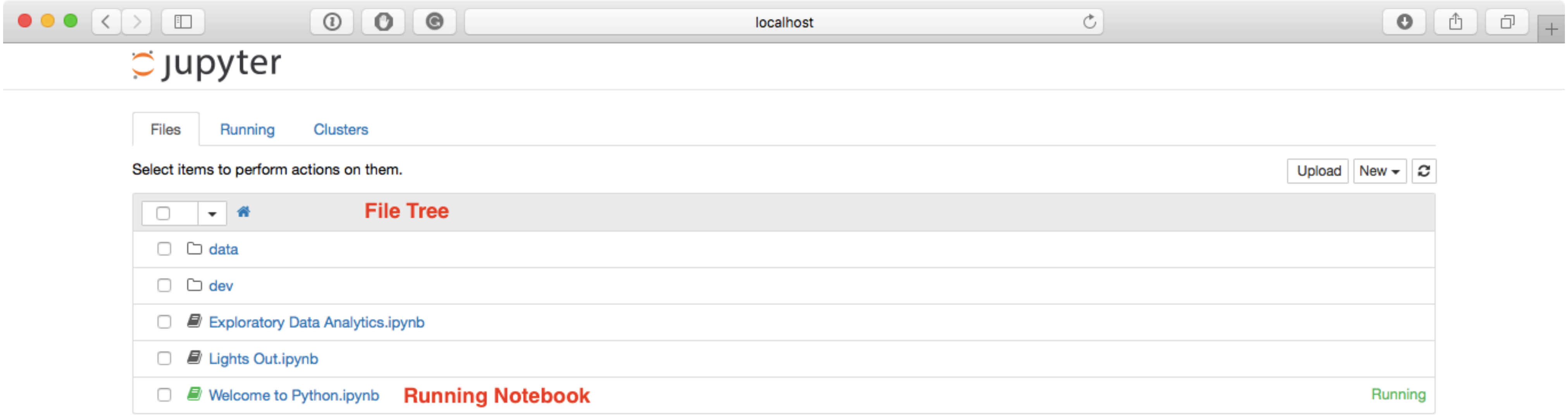
- Clustergrammer
    - http://amp.pharm.mssm.edu/clustergrammer/

What is Anaconda and Jupyter notebook?

- Python, Jupyter notebook and Anaconda
  - https://www.anaconda.com/products/individual
  - https://github.com/MaayanLab/clustergrammer-widget

- Jupyter
  https://jupyter-notebook.readthedocs.io/en/stable/ui_components.html

What is Anaconda and Jupyter notebook?



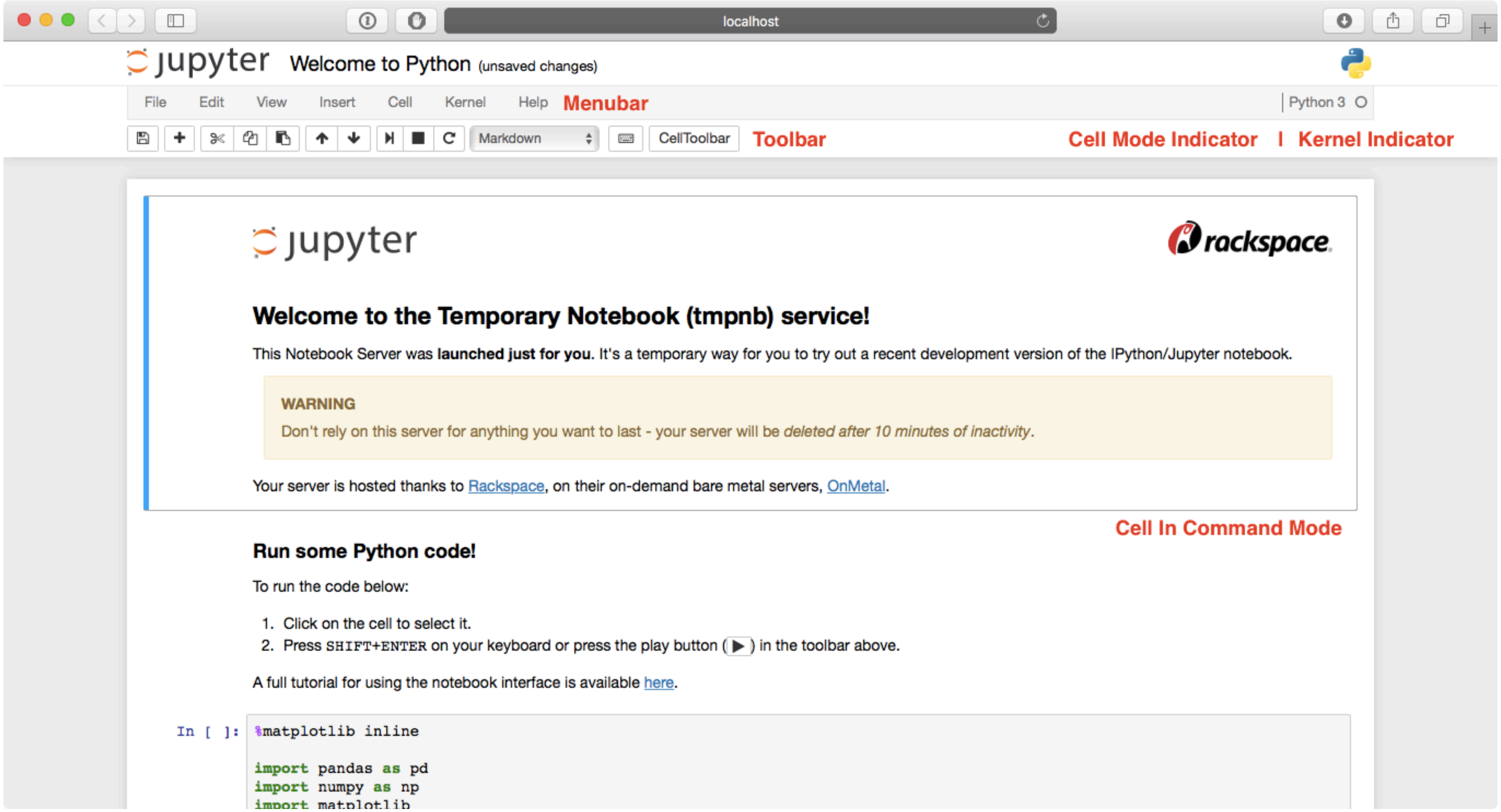
Select items to perform actions on them.

File Tree

- □ data
- □ dev
- □ Exploratory Data Analytics.ipynb
- □ Lights Out.ipynb
- □ Welcome to Python.ipynb Running

Running Notebook

What is Anaconda and Jupyter notebook?


localhost
jupyter Welcome to Python (unsaved changes)
File Edit View Insert Cell Kernel Help
Menubar
Python 3
Markdown
CellToolbar
Toolbar
Cell Mode Indicator | Kernel Indicator
jupyter
rackspace
Cell In Command Mode

Welcome to the Temporary Notebook (tmpnb) service!

This Notebook Server was **launched just for you**. It's a temporary way for you to try out a recent development version of the IPython/Jupyter notebook.

WARNING
Don't rely on this server for anything you want to last - your server will be *deleted after 10 minutes of inactivity*.

Your server is hosted thanks to Rackspace, on their on-demand bare metal servers, OnMetal.

Run some Python code!

To run the code below:

1. Click on the cell to select it.
2. Press `SHIFT+ENTER` on your keyboard or press the play button (▶) in the toolbar above.

A full tutorial for using the notebook interface is available here.

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Setup conda

```
# open terminal
# if your Anaconda is installed you should have “(base)” prefix in your terminal prompt

conda create -n mutations python=3.7 jupyter
conda activate mutations

pip install mutagene
pip install pyyaml
pip install pandas==0.25.1

pip install clustergrammer-widget
jupyter nbextension enable --py --sys-prefix widgetsnbextension
jupyter nbextension enable --py --sys-prefix clustergrammer_widget

mutagene --version

# change directory to where you will be analyzing data and storing files
jupyter notebook
```

Setup mutagene package

```
# change directory to where you will be analyzing data and storing files
# choose appropriate genome assembly
# -g hg38 for GRCh38
# -g hg19 for GRCh37

mutagene fetch genome -g hg38
```

Run MutaGene to identify mutational signatures

```
mutagene identify
        -g hg38
        -i TCGA.COAD.mutect.03652df4-6090-4f5a-a2ff-ee28a37f9301.DR-10.0.somatic.maf
        -o COAD.txt
        -f MAF
        -s COSMICv3
```

Select genome assembly (hg38)
Specify input file and output file
Specify output file format (MAF)
Select set of signatures (COSMICv3)

```
# for other options see
mutagene identify --help
```

```
# mutational decomposition into signatures results are now in COAD.txt file
(mutations) HG-02113788-LM4:MutationalAnalysisDemo gonceare$ head COAD.txt
sample     signature  exposure   mutations
TCGA-AA-3966-01A-01D-1981-10     COSMICv3-SBS15       0.223284   201
TCGA-AA-3966-01A-01D-1981-10     COSMICv3-SBS6        0.125008   113
TCGA-AA-3966-01A-01D-1981-10     COSMICv3-SBS20       0.114956   104
TCGA-AA-3966-01A-01D-1981-10     COSMICv3-SBS54       0.0984864  89
…
```

Run jupyter notebook

```
# to run notebook server and open a browser window:
jupyter notebook
```

https://github.com/elnitskilab/MutationalAnalysisDemo