Josh Graham
TGF Lab – Lehigh University

CRISPR-GEM Tutorial
January 14th, 2025

CRISPR-GEM v1 Tutorial

Table of Contents:

Josh Graham
TGF Lab – Lehigh University

CRISPR-GEM Tutorial
January 14th, 2025

Overview:

CRISPR-GEM is a machine learning model built to identify promising CRISPR target genes for specific gene editing applications. There are two individual components of CRISPR-GEM, a graphical user interface (GUI) and a deep neural network (DNN) model in Google Colab. The first is a GUI that is used to compare different cell types and narrow the list of genes to more precisely model CRISPR strategies. The GUI provides a list of input genes and a list of output genes that are candidate genes for the editing strategy or important markers to assess a successful therapy respectively. To accomplish this, users must specify the tissue/cell types to be edited (ex. Diseased Cartilage), the goal for gene editing (ex. Healthy Cartilage), and the type of CRISPR gene editing to be used (ex. CRISPRa) using the GUI.

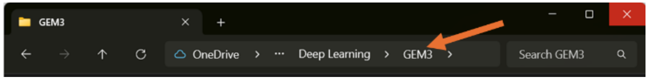
Once the appropriate cell/tissue types are selected and the GUI is used to obtain the input/output genes for the model, this information can be transferred over to a google drive account for the assessment of each gene candidate. The DNN learns the expression of the output genes provided by the GUI using the expression of the input genes then perturbs input gene expression in a CRISPR-mimetic fashion to assess the downstream effect on output gene expression. The result is a ranked list of CRISPR candidates that score the best for accomplishing a therapeutic goal using CRISPR. These genes should be assessed *in vitro* to verify their accuracy.

It is important to note that this model is highly sensitive to the datasets used to identify gene candidates in the GUI and those fed to the DNN model. When comparing sequencing datasets, it is advised to always use data from the same experiment when possible and at the very least from the same sequencing platform. The DNN is trained only on HiSeq2000 and HiSeq2500 data to reduce any noise resulting from different platforms. If users have their own datasets for these platforms they can be easily integrated into the existing datasets and used for analysis (pg. 10). It is also possible for users to select new datasets for analysis through the Gene Expression Omnibus (GEO) to add to existing datasets (pg. 2). Additionally, users may select target genes manually via differential expression analysis or alternative methods if they have data that is not yet public. Users must be reminded that when obtaining new datasets or using their own, meaningful integration into the model requires that they are sequenced using HiSeq2000 and HiSeq2500. However, adaptations are underway to accommodate alternative datasets.

Downloading GUI:

- Download python 3.11.5
    - Python 3.11.5 is required since some modules have been depreciated in newer versions.
    - Make sure that you add it to your systems PATH
    - If a later version of python is downloaded it may be necessary to use a virtual environment to run python 3.11.1.
- The datasets for the GUI & model training are different. They will both have to be personally shared over google drive so email Josh at jog822@lehigh.edu for access.
- The remaining components for the CRISPR-GEM can be downloaded from the CRISPR-GEM Github.
- Download the Github files and dataset to the same location then open a command prompt. (On windows type cmd into the search bar and immediately hit enter).
- Identify the file location of where you downloaded the files.

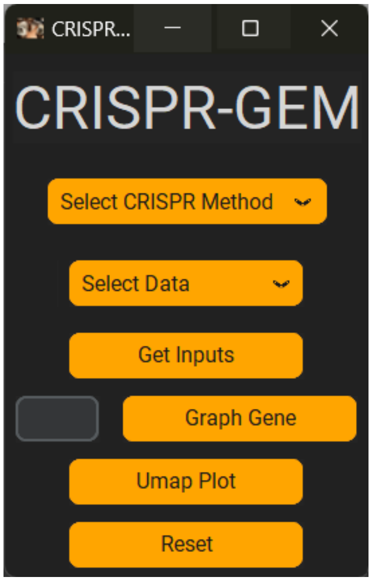
  Ex.

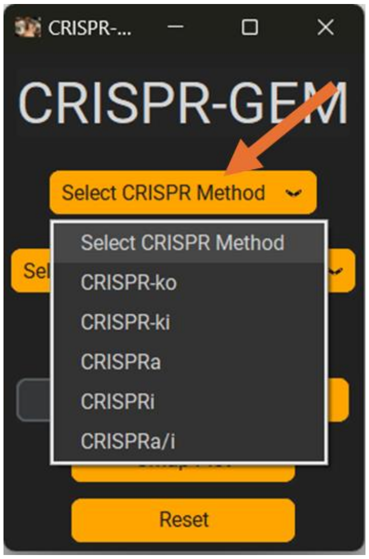
  Right click on the file location and click “Copy Address as Text”.
- In the command window type “cd ” then paste the address and hit enter.
- Now you can run the script by typing “python GUI.py”

Using GUI:

1. Copy the folder location that GUI files and datasets are in
2. In the command window type "cd " then paste the address and hit enter.
3. Now you can run the script by typing "python GUI.py"
4. The GUI will take 1-2 minutes to load and should look like this:

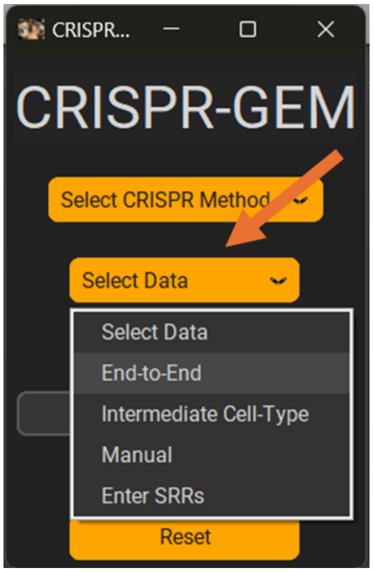


5. Start by selecting the CRISPR strategy based on if you want to perform gene knock-in, knock-out, activation, or repression.

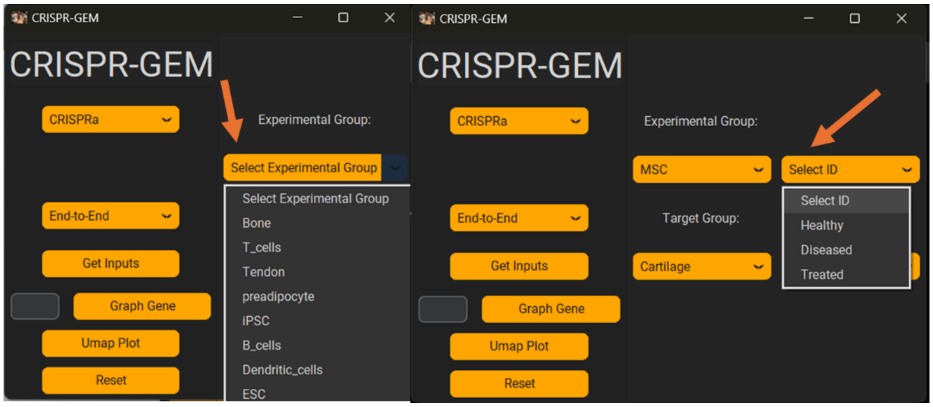


6. Then select the format with which you want to set up the model (Select Data). This has to do with how many cell/tissue types that you want to enter. End-to-End (shown here) uses just two cell tissue types. However, the user can also choose to add a third type using the "Intermediate Cell Type" option. This gives the model a trajectory to work off of. For example, when analyzing chondrogenesis, the Intermediate cell type for MSC chondrogenesis was fetal cartilage which represents the natural

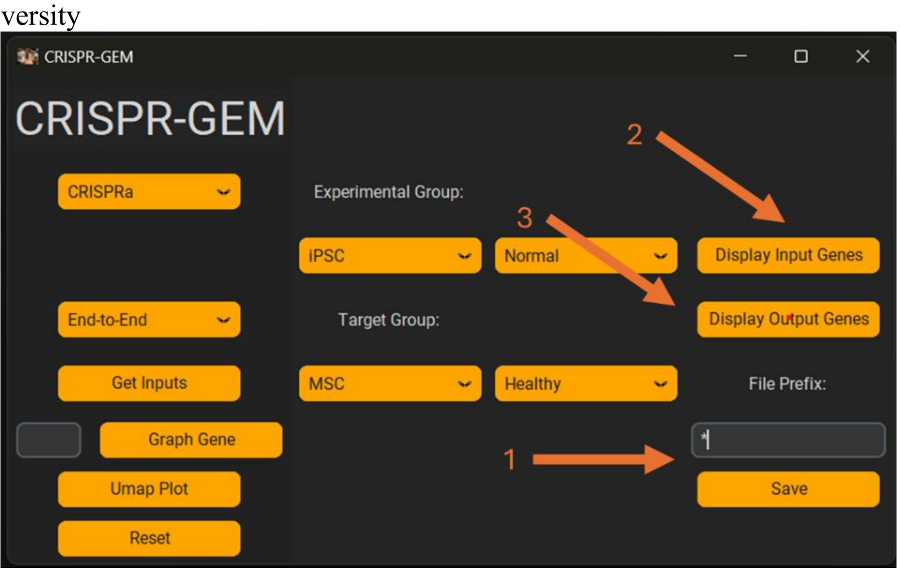
development pathway. Alternatively, the manual method can be used to manually enter genes by giving their Ensemble IDs.



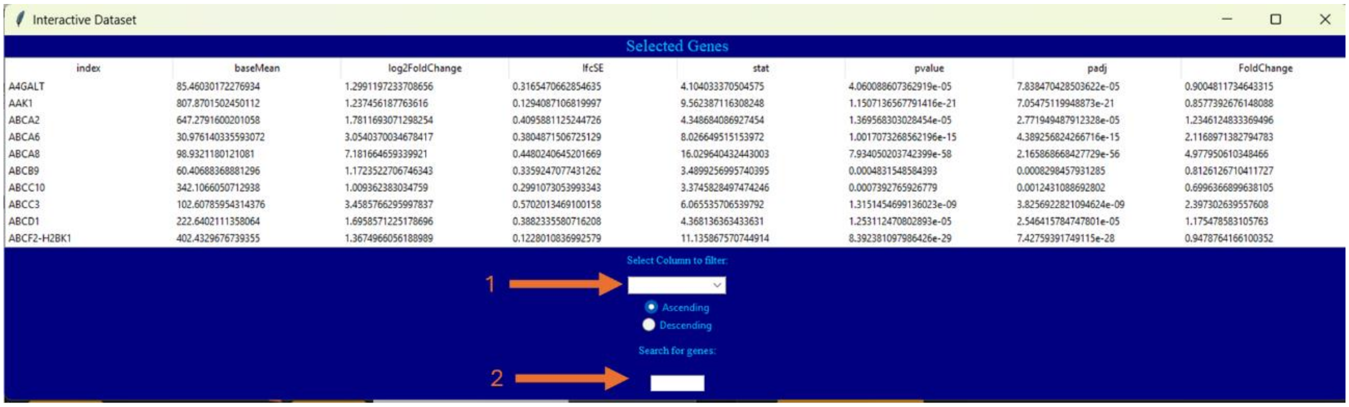
7. The actual cell/tissue types are then selected using two categories. The first is a broad description of the group (ex. MSC, chondrocyte, fibroblast) is labeled ‘Select Experimental Group’ and pops up first. Then a more specific description such as healthy, diseased, treated, etc. can be selected in the ‘Select ID’ dropdown that will pop up to the right of the ‘Select Experimental Group’ tab.



8. Once everything is selected, one can then select for the input/output genes based on those groups using the ‘Get Inputs’ button.
9. This will take a couple minutes to run but once it is finished a new window will pop up in the corner with an entry box and ‘Save’ button (1, below). The user should type a short tag that describes the experiment here and save the file. This will create a new folder in the directory. For example, let * represent the short tag, then the output will be files will be in a folder named ‘*’ in the GUI directory with files named ‘*_input_list.csv’, ‘*_output_list.csv’, ‘*_SRRs_used.csv’, ‘*_input_genes.csv’, and ‘*_output_genes’.



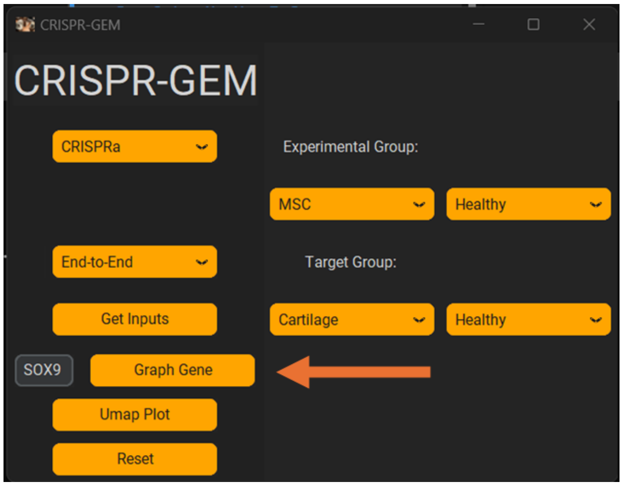
10. The results can be monitored by clicking the ‘Display Input Genes’ or ‘Display Output Genes’ buttons (2 & 3 above respectively). These give interactive datasets that you can use to view the results. This can be done by selecting a column (ex. Log2FoldChange) and sorting it by ascending or descending value (1, below). Or a user can input a specific gene (make sure it is all caps) in the entry box (2, below) to find well established genes.for comparison.


Interactive Dataset

Selected Genes

index	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	FoldChange
A4GALT	85.46030172276934	1.2991197233708656	0.3165470662854635	4.104033370504575	4.060088607362919e-05	7.838470428503622e-05	0.9004811734643315
AAK1	807.8701502450112	1.237456187763616	0.1294087106819997	9.562387116308248	1.1507136567791416e-21	7.05475119948873e-21	0.8577392676148088
ABCA2	647.2791600201058	1.7811693071298254	0.4095881125244726	4.348684086927454	1.369568303028454e-05	2.771949487912328e-05	1.2346124833369496
ABCA6	30.976140335593072	3.0540370034678417	0.3804871506725129	8.026649515153972	1.0017073268562196e-15	4.389256824266716e-15	2.1168971382794783
ABCA8	98.9321180121081	7.181664659339921	0.4480240645201669	16.029640432443003	7.934050203742399e-58	2.165868668427729e-56	4.977950610348466
ABCB9	60.40688368881296	1.1723522706746343	0.3359247077431262	3.4899256995740395	0.0004831548584393	0.0008298457931285	0.8126126710411727
ABCC10	342.1066050712938	1.009362383034759	0.2991073053993343	3.3745828497474246	0.0007392765926779	0.0012431088692802	0.6996366899638105
ABCC3	102.60785954314376	3.4585766295997837	0.5702013469100158	6.065535706539792	1.3151454699136023e-09	3.8256922821094624e-09	2.397302639557608
ABCD1	222.6402111358064	1.6958571225178696	0.3882335580716208	4.368136363433631	1.253112470802893e-05	2.546415784747801e-05	1.175478583105763
ABCF2-H2BK1	402.4329676739355	1.3674966056188989	0.1228010836992579	11.135867570744914	8.392381097986426e-29	7.42759391749115e-28	0.9478764166100352

11. To evaluate datasets before using the identified input/output genes, one can graph the relative expression of a specific gene by typing it into the entry box then clicking the “Graph Gene” button next to it (ex. *SOX9* below). The plots the relative expression in each sample.



CRISPRa

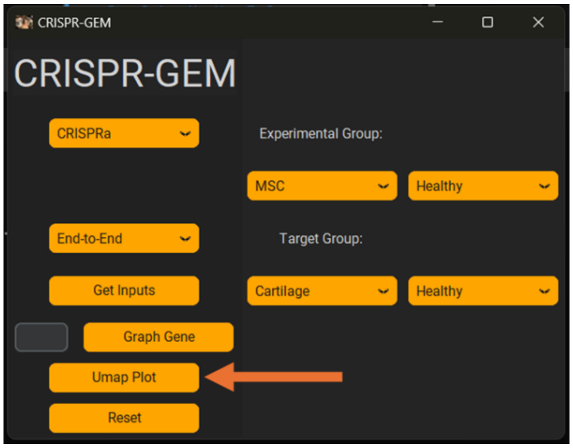
End-to-End

Experimental Group: MSC, Healthy

Target Group: Cartilage, Healthy

SOX9

12. Additionally, more global representations of the datasets for each selected cell/tissue type can be represented by clicking the ‘Umap Plot’ button.



CRISPRa

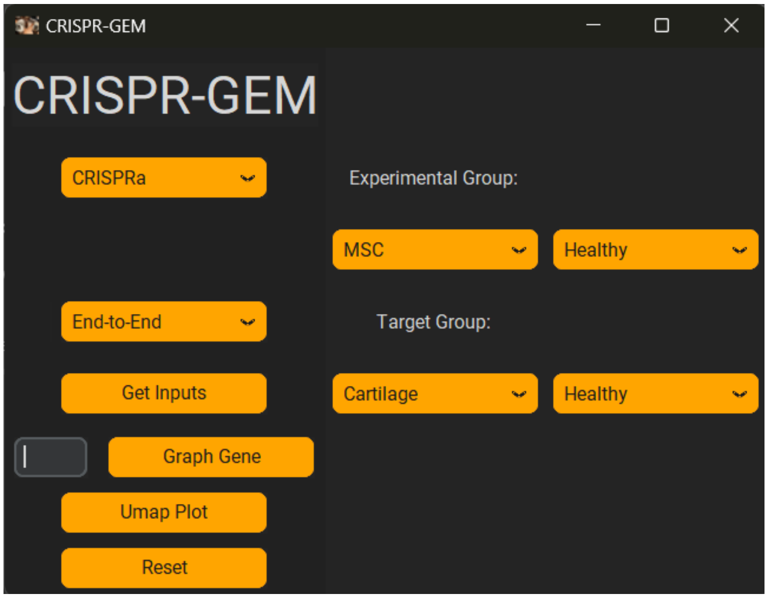
End-to-End

Experimental Group: MSC, Healthy

Target Group: Cartilage, Healthy

Troubleshooting Misaligned Data:

Ex. Healthy cartilage and healthy MSCs using the end-to-end method



CRISPR-GEM

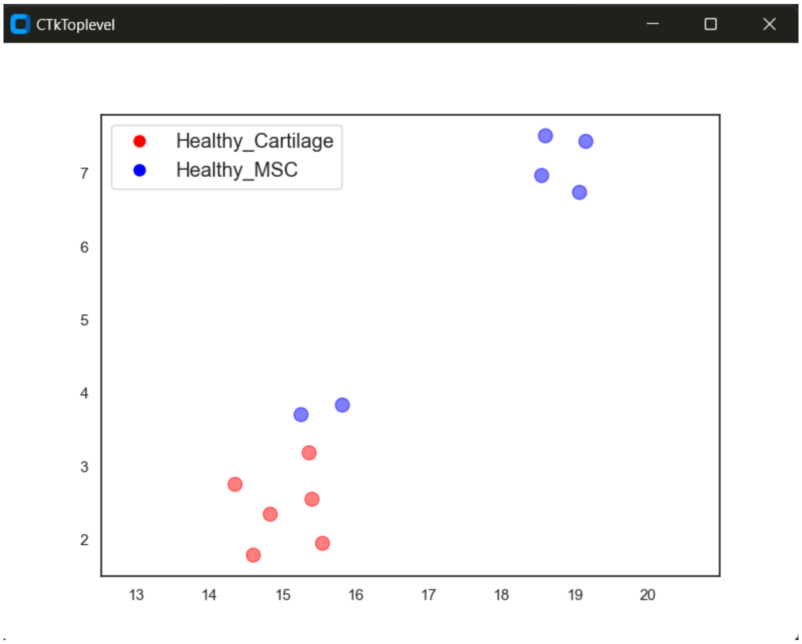
CRISPRa

Experimental Group: MSC Healthy

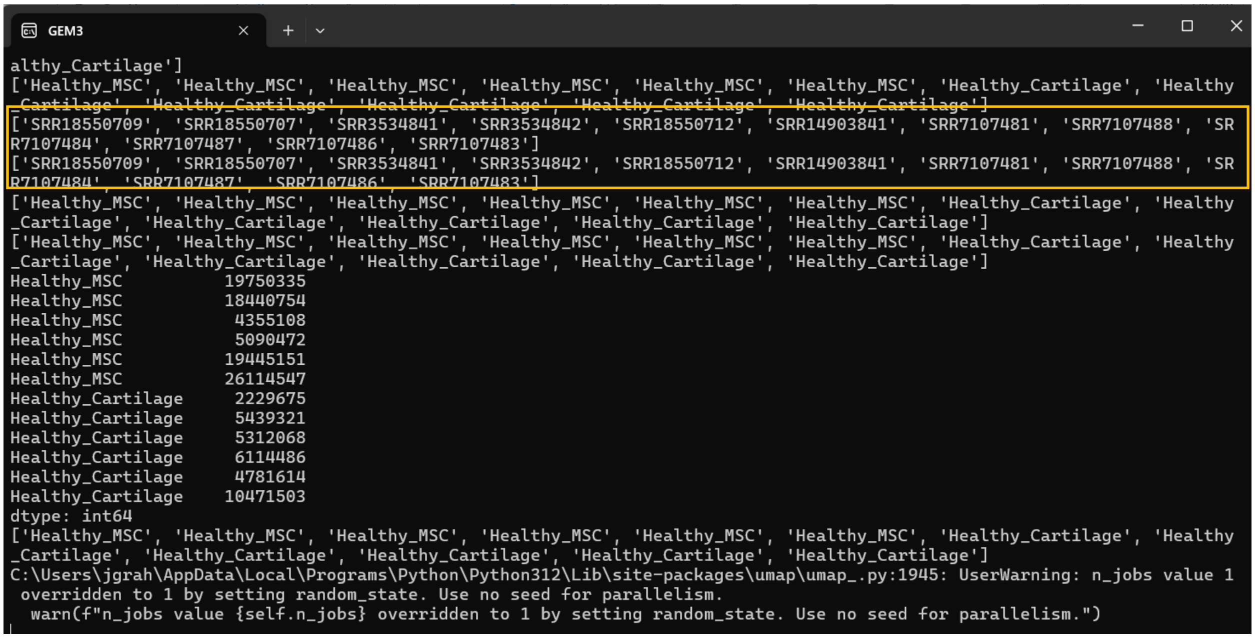
End-to-End

Target Group: Cartilage Healthy

But before pressing the “Get Inputs” button, it is advised to evaluate the underlying data to ensure that the input and output genes are appropriately assessed. To do so, one can start by using the UMAP function to visualize the data in a reduced dimension to get an idea of how similar the data from each group is. In doing so, the following graph is plotted:

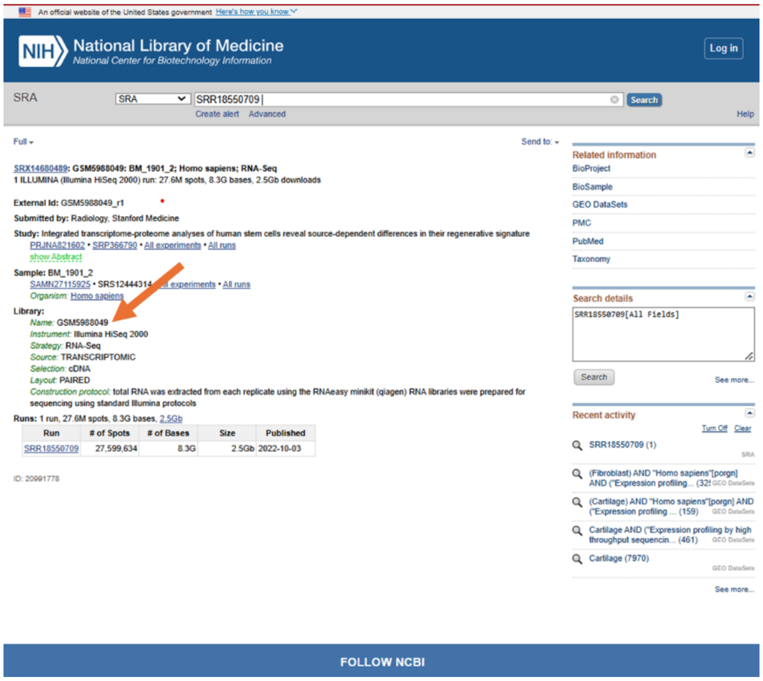


There are two clear clusters for each group, however two of the six MSC datapoints seem to be more similar to the Cartilage group which may interfere with differential expression analysis. To overcome this, one can look at the datasets used for analysis using the underlying command window which should look something like this: .



```
althy_Cartilage']
['Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_Cartilage', 'Healthy
_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage']
['SRR18550709', 'SRR18550707', 'SRR3534841', 'SRR3534842', 'SRR18550712', 'SRR14903841', 'SRR7107481', 'SRR7107488', 'SR
R7107484', 'SRR7107487', 'SRR7107486', 'SRR7107483']
['SRR18550709', 'SRR18550707', 'SRR3534841', 'SRR3534842', 'SRR18550712', 'SRR14903841', 'SRR7107481', 'SRR7107488', 'SR
R7107484', 'SRR7107487', 'SRR7107486', 'SRR7107483']
['Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_Cartilage', 'Healthy
_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage']
['Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_Cartilage', 'Healthy
_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage']
Healthy_MSC          19750335
Healthy_MSC          18440754
Healthy_MSC           4355108
Healthy_MSC           5090472
Healthy_MSC          19445151
Healthy_MSC          26114547
Healthy_Cartilage     2229675
Healthy_Cartilage     5439321
Healthy_Cartilage     5312068
Healthy_Cartilage     6114486
Healthy_Cartilage     4781614
Healthy_Cartilage    10471503
dtype: int64
['Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_MSC', 'Healthy_Cartilage', 'Healthy
_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage', 'Healthy_Cartilage']
C:\Users\jgrah\AppData\Local\Programs\Python\Python312\Lib\site-packages\umap\umap_.py:1945: UserWarning: n_jobs value 1
 overridden to 1 by setting random_state. Use no seed for parallelism.
  warn(f"n_jobs value {self.n_jobs} overridden to 1 by setting random_state. Use no seed for parallelism.")
```

The datasets used can be identified in the given lists and searched in the SRA for more insight. It is often helpful to find the GEO ID then search the GEO for the dataset as it provides more information.


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
SRA
SRA
SRR18550709
Search
Create alert
Advanced
Help
Full
Send to:
Related information
BioProject
BioSample
GEO DataSets
PMC
PubMed
Taxonomy
Search details
SRR18550709[All Fields]
Search
See more...
Recent activity
Turn Off
Clear
SRR18550709 (1) SRA
(Fibroblast) AND "Homo sapiens"[porgn] AND ("Expression profiling... (32 GEO DataSets
(Cartilage) AND "Homo sapiens"[porgn] AND ("Expression profiling ... (159) GEO DataSets
Cartilage AND ("Expression profiling by high throughput sequencin... (461) GEO DataSets
Cartilage (7970) GEO DataSets
See more...
FOLLOW NCBI

SRX14680489: GSM5988049: BM_1901_2; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 27.6M spots, 8.3G bases, 2.5Gb downloads

External Id: GSM5988049_r1

Submitted by: Radiology, Stanford Medicine

Study: Integrated transcriptome-proteome analyses of human stem cells reveal source-dependent differences in their regenerative signature
PRJNA821602 • SRP366790 • All experiments • All runs
show Abstract

Sample: BM_1901_2
SAMN27115925 • SRS12444314 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: GSM5988049
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED
Construction protocol: total RNA was extracted from each replicate using the RNAeasy minikit (qiagen) RNA libraries were prepared for sequencing using standard Illumina protocols

Runs: 1 run, 27.6M spots, 8.3G bases, 2.5Gb

Run	# of Spots	# of Bases	Size	Published
SRR18550709	27,599,634	8.3G	2.5Gb	2022-10-03

ID: 20991778

Once the misaligned dataset is found, the inappropriate datasets can be found in the 'Labels.csv' file and removed by changing the tag by adding a number to the end or by adding a better description. For example, a likely change source of error in the given circumstance is that MSCs were senescent, differentiated, or harvested from an older source so the description 'Healthy' could be changed to 'Senescent' or just 'Healthy1' to remove it from the analysis. Don't forget to change that label back to its original description if it was not the cause of the issue.

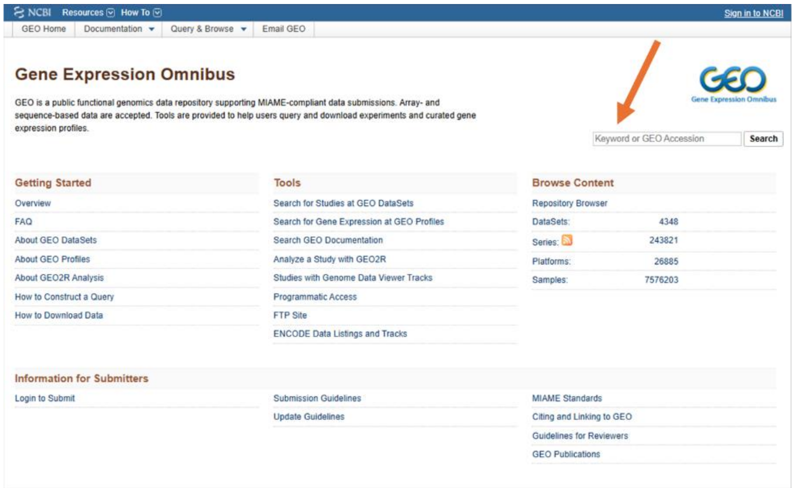
Sometimes it is hard to notice any significant differences just from the descriptions on the SRA or GEO, hence why they were all labeled the same. An alternative option is to eliminate specific datasets one-by-one to monitor which is misaligned. This can be done in two ways. **1)** Use the 'Enter SRRs' option in the "Select Datasets" dropdown menu. This allows you to paste a comma separated list (ex. SRR123, SRR124, …) so it is easy to eliminate a dataset or group of similar datasets one-by-one to identify the inappropriate dataset. **2)** Changing items in the "Labels.csv" file to identify the issue. It is advised to change all datasets containing the same 'GEO' accession (ex. GSE1234…) at once since these are most likely collected at the same time and from the same source. In both cases, once the changes are made, the data can be plotted using the UMAP button again to view the effects. Don't forget to change that label back to its original description if it was not the cause of the issue.

Please share any inappropriately labeled datasets with jog822@lehigh.edu. In addition to using UMAP. Users can assess datasets by graphing multiple gene, although this strategy is not as encompassing as UMAP.

Adding new datasets into the CRISPR-GEM Database:

Data may need to be selected in the case that the appropriate datasets are not available on the GUI or if there is not a deep enough data n for that specific cell/tissue type on the GUI.

- All data can be selected from the gene expression omnibus: https://www.ncbi.nlm.nih.gov/geo/.
- Search for appropriate datasets using the search bar.

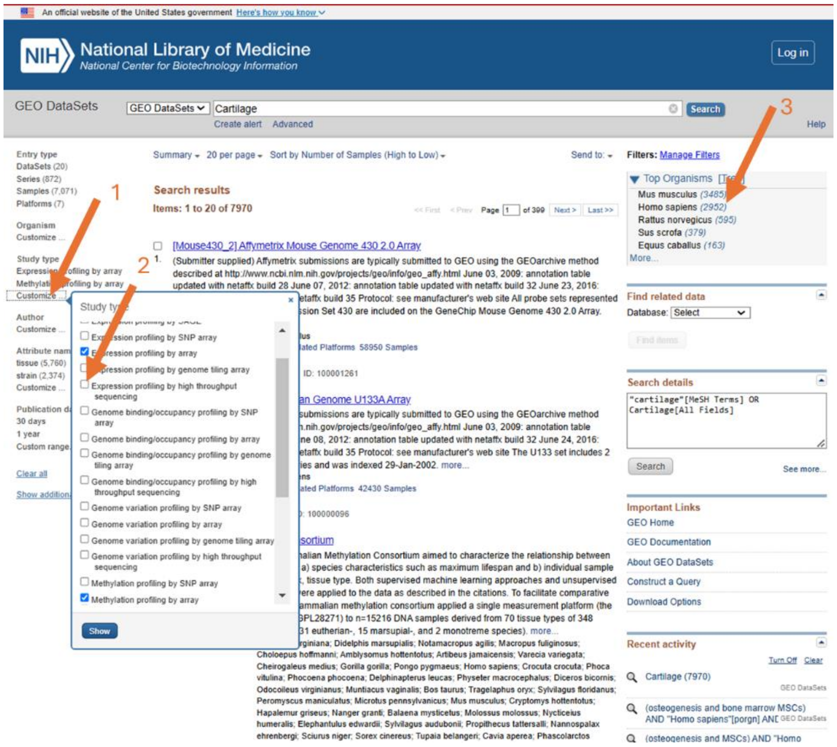


Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

DataSets:	4348
Series:	243821
Platforms:	26885
Samples:	7576203

- In the search results make sure that you filter results for the appropriate data types. Click 'Customize' under Study Type (1), then make sure 'Expression profiling through High Throughput Sequencing' is checked (2). This option will then appear under Study Type. Click 'Expression profiling through High Throughput Sequencing' again to filter results. Then, filter for only human datasets by clicking the number next to 'Homo Sapiens' (3).


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
GEO DataSets
GEO DataSets
Cartilage
Search
Create alert
Advanced
Help
3
Summary
20 per page
Sort by Number of Samples (High to Low)
Send to:
Filters: Manage Filters
Entry type
DataSets (20)
Series (872)
Samples (7,071)
Platforms (7)
Organism
Customize ...
Study type
Expression profiling by array
Methylation profiling by array
Customize ...
Author
Customize ...
Attribute name
tissue (5,760)
strain (2,374)
Customize ...
Publication date
30 days
1 year
Custom range...
Clear all
Show additional
1
2
Study type
Expression profiling by SAGE
☐ Expression profiling by SNP array
☑ Expression profiling by array
☐ Expression profiling by genome tiling array
☐ Expression profiling by high throughput sequencing
☐ Genome binding/occupancy profiling by SNP array
☐ Genome binding/occupancy profiling by array
☐ Genome binding/occupancy profiling by genome tiling array
☐ Genome binding/occupancy profiling by high throughput sequencing
☐ Genome variation profiling by SNP array
☐ Genome variation profiling by array
☐ Genome variation profiling by genome tiling array
☐ Genome variation profiling by high throughput sequencing
☐ Methylation profiling by SNP array
☑ Methylation profiling by array
Show
Top Organisms [Tree]
Mus musculus (3485)
Homo sapiens (2952)
Rattus norvegicus (595)
Sus scrofa (379)
Equus caballus (163)
More...
Find related data
Database: Select
Find items
Search details
Search
See more...
Important Links
GEO Home
GEO Documentation
About GEO DataSets
Construct a Query
Download Options
Recent activity
Turn Off
Clear
Cartilage (7970)
GEO DataSets
(osteogenesis and bone marrow MSCs) AND "Homo sapiens"[porgn] AND
GEO DataSets
(osteogenesis and MSCs) AND "Homo

Search results

Items: 1 to 20 of 7970

<< First < Prev Page 1 of 399 Next > Last >>

☐ [Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array

1. (Submitter supplied) Affymetrix submissions are typically submitted to GEO using the GEOarchive method described at http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_affy.html June 03, 2009: annotation table updated with netaffx build 28 June 07, 2012: annotation table updated with netaffx build 32 June 23, 2016: annotation table updated with netaffx build 35 Protocol: see manufacturer's web site All probe sets represented on the GeneChip Mouse Expression Set 430 are included on the GeneChip Mouse Genome 430 2.0 Array.

Mus musculus
Related Platforms 58950 Samples

ID: 100001261

an Genome U133A Array

submissions are typically submitted to GEO using the GEOarchive method
.nih.gov/projects/geo/info/geo_affy.html June 03, 2009: annotation table
ne 08, 2012: annotation table updated with netaffx build 32 June 24, 2016:
etaffx build 35 Protocol: see manufacturer's web site The U133 set includes 2
ies and was indexed 29-Jan-2002. more...

ns
ated Platforms 42430 Samples

: 100000096

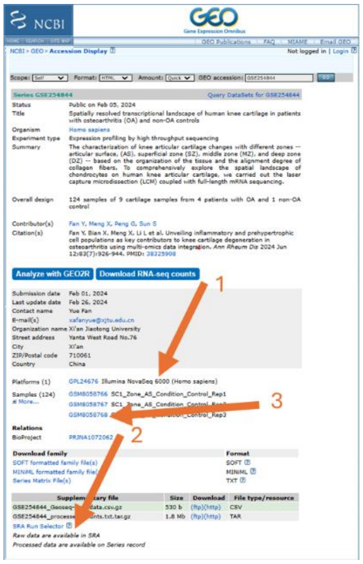
sortium

alian Methylation Consortium aimed to characterize the relationship between a) species characteristics such as maximum lifespan and b) individual sample , tissue type. Both supervised machine learning approaches and unsupervised ere applied to the data as described in the citations. To facilitate comparative ammalian methylation consortium applied a single measurement platform (the GPL28271) to n=15216 DNA samples derived from 70 tissue types of 348 31 eutherian-, 15 marsupial-, and 2 monotreme species). more...

rginiana; Didelphis marsupialis; Notamacropus agilis; Macropus fuliginosus; Choloepus hoffmanni; Amblysomus hottentotus; Artibeus jamaicensis; Varecia variegata; Cheirogaleus medius; Gorilla gorilla; Pongo pygmaeus; Homo sapiens; Crocuta crocuta; Phoca vitulina; Phocoena phocoena; Delphinapterus leucas; Physeter macrocephalus; Diceros bicornis; Odocoileus virginianus; Muntiacus vaginalis; Bos taurus; Tragelaphus oryx; Sylvilagus floridanus; Peromyscus maniculatus; Microtus pennsylvanicus; Mus musculus; Cryptomys hottentotus; Hapalemur griseus; Nanger granti; Balaena mysticetus; Molossus molossus; Nycticeius humeralis; Elephantulus edwardii; Sylvilagus audubonii; Propithecus tattersalli; Nannospalax ehrenbergi; Sciurus niger; Sorex cinereus; Tupaia belangeri; Cavia aperea; Phascolarctos

```
"cartilage"[MeSH Terms] OR
Cartilage[All Fields]
```

- All Data sets used for model training should be taken from HiSeq2000 or HiSeq2500; however, for simply determining the input/output genes for model training other platforms can be used as long as differential expression is always performed across datasets from the same platform. The platform is generally listed on the main page for a dataset (1, below). However, in some cases multiple platforms are listed, and one must click on individual datapoints to identify the platform.



Series GSE254844   Query DataSets for GSE254844

Status	Public on Feb 05, 2024
Title	Spatially resolved transcriptional landscape of human knee cartilage in patients with osteoarthritis (OA) and non-OA controls
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The characterization of knee articular cartilage changes with different zones -- articular surface, (AS), superficial zone (SZ), middle zone (MZ), and deep zone (DZ) -- based on the organization of the tissue and the alignment degree of collagen fibers. To comprehensively explore the spatial landscape of chondrocytes on human knee articular cartilage, we carried out the laser capture microdissection (LCM) coupled with full-length mRNA sequencing.
Overall design	124 samples of 9 cartilage samples from 4 patients with OA and 1 non-OA control
Contributor(s)	Fan Y, Meng X, Peng G, Sun S
Citation(s)	Fan Y, Bian X, Meng X, Li L et al. Unveiling inflammatory and prehypertrophic cell populations as key contributors to knee cartilage degeneration in osteoarthritis using multi-omics data integration. *Ann Rheum Dis* 2024 Jun 12;83(7):926-944. PMID: 38325908

Submission date	Feb 01, 2024
Last update date	Feb 26, 2024
Contact name	Yue Fan
E-mail(s)	xafanyue@xjtu.edu.cn
Organization name	Xi'an Jiaotong University
Street address	Yanta West Road No.76
City	Xi'an
ZIP/Postal code	710061
Country	China

Platforms (1)	GPL24676 Illumina NovaSeq 6000 (Homo sapiens)
Samples (124) More...	GSM8058766 SC1_Zone_AS_Condition_Control_Rep1 GSM8058767 SC1_Zone_AS_Condition_Control_Rep2 GSM8058768 SC1_Zone_AS_Condition_Control_Rep3

Relations

BioProject	PRJNA1072062

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

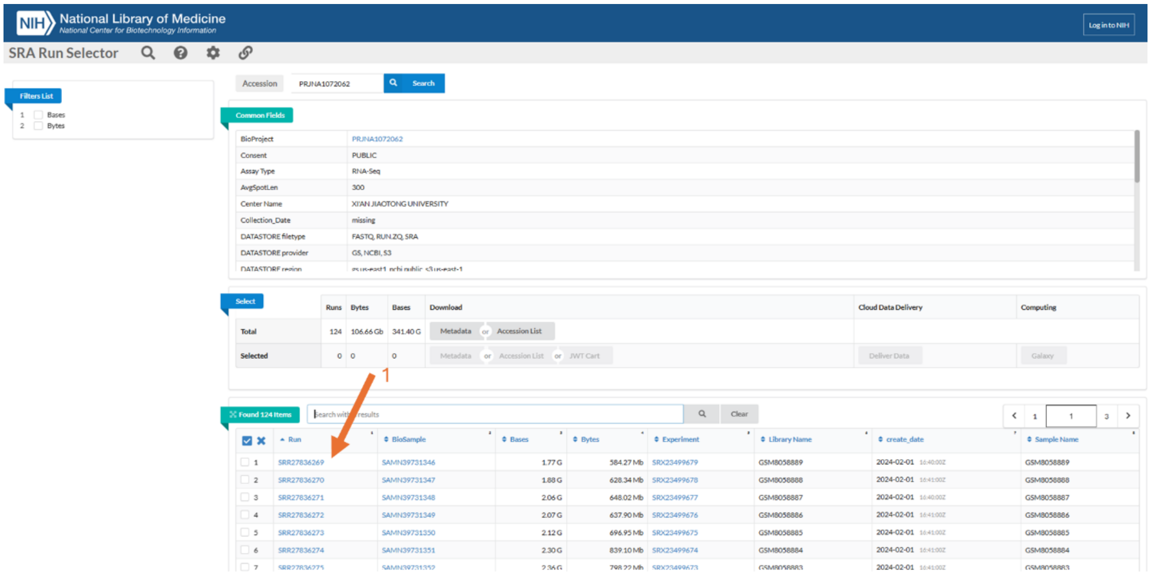
Supplementary file	Size	Download	File type/resource
GSE254844_Geoseq-[illegible]data.csv.gz	530 b	(ftp)(http)	CSV
GSE254844_processe[illegible]nts.txt.tar.gz	1.8 Mb	(ftp)(http)	TAR

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

- To download the data click on SRA selector (2, above). The data must be downloadable from SRA for each addition to datasets. All that is needed is the SRA number for each dataset which can be linked to the GSM accession number in the dataset page (3, above).



Accession: PRJNA1072062

Filters List
1 □ Bases
2 □ Bytes

Common Fields

BioProject	PRJNA1072062
Consent	PUBLIC
Assay Type	RNA-Seq
AvgSpotLen	300
Center Name	XI'AN JIAOTONG UNIVERSITY
Collection_Date	missing
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.us-east1, ncbi.public, s3.us-east-1

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	124	106.66 Gb	341.40 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 124 Items

Search within results

‹ 1 [1] 3 ›

☑ ✖	Run	BioSample	Bases	Bytes	Experiment	Library Name	create_date	Sample Name
□ 1	SRR27836269	SAMN39731346	1.77 G	584.27 Mb	SRX23499679	GSM8058889	2024-02-01 16:40:00Z	GSM8058889
□ 2	SRR27836270	SAMN39731347	1.88 G	628.34 Mb	SRX23499678	GSM8058888	2024-02-01 16:41:00Z	GSM8058888
□ 3	SRR27836271	SAMN39731348	2.06 G	648.02 Mb	SRX23499677	GSM8058887	2024-02-01 16:40:00Z	GSM8058887
□ 4	SRR27836272	SAMN39731349	2.07 G	637.90 Mb	SRX23499676	GSM8058886	2024-02-01 16:41:00Z	GSM8058886
□ 5	SRR27836273	SAMN39731350	2.12 G	696.95 Mb	SRX23499675	GSM8058885	2024-02-01 16:41:00Z	GSM8058885
□ 6	SRR27836274	SAMN39731351	2.30 G	839.10 Mb	SRX23499674	GSM8058884	2024-02-01 16:41:00Z	GSM8058884
□ 7	SRR27836275	SAMN39731352	2.36 G	798.22 Mb	SRX23499673	GSM8058883	2024-02-01 16:41:00Z	GSM8058883

- These SRA numbers may be either added to an excel file/CSV formatted as follows. Or copied into a comma separated list eg. ‘SRA1’, ‘SRA2’, ….

SRR ID
SRR1
SRR2

** Any other information can follow such as the GSM #, cell type, disease state, etc.

- Run the code “Kallisto_sra_downloads_Official.ipynb” with the corresponding SRR codes (See line 34 or 36/37). Make sure to follow the prompts.
- To add this data to the dataset for the GUI, add ‘Big.csv’ to Google Drive in same place as TPM.csv. Then toggle (add and remove ‘#’) back and forth between the pairs of lines below, with the bottom always representing the “Big.csv” and the top line representing “TPM.csv”.

```
hello = pd.read_csv('/content/drive/MyDrive/TIGR_/TPM.csv', index_col=0)
# hello = pd.read_csv('/content/drive/MyDrive/TIGR_/Big.csv', index_col=0)
```

```
est_counts = tsv_df['tpm'].rename(folder)
# est_counts = tsv_df['est_counts'].rename(folder)
```

```
TPM.to_csv('/content/drive/MyDrive/TIGR_/TPM.csv')
#TPM.to_csv('/content/drive/MyDrive/TIGR_/Big.csv')
```

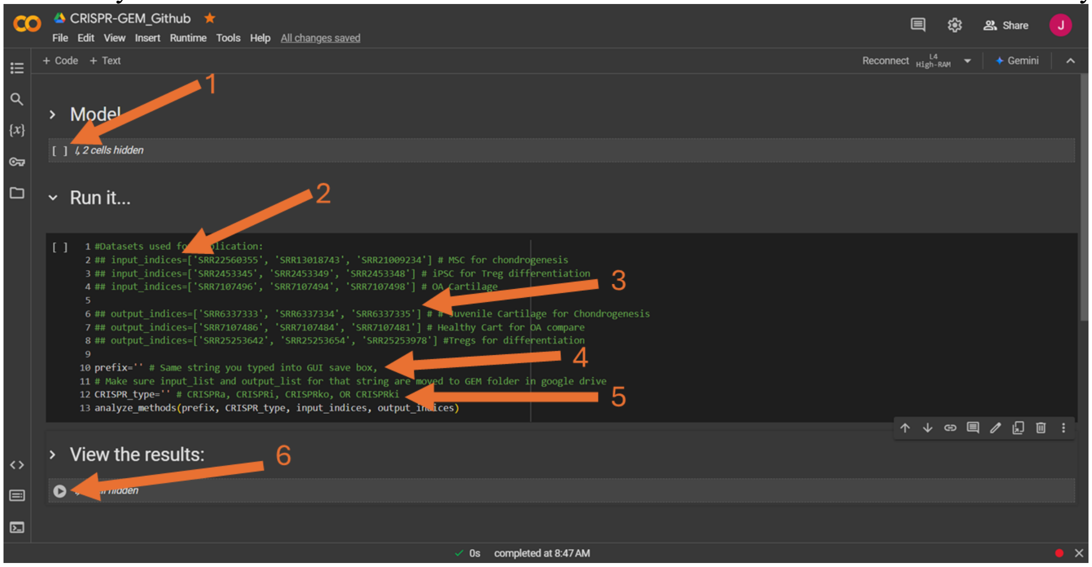
Using CRISPR-GEM in Google Colab.

1. Create a New Folder in Google Drive (Your ‘My Drive’) Called “GEM”.
2. Add the *_input_list and *_output_list into this folder from the GUI.

   * It is also possible to create your own list of genes for each group without the use of the GEO. For example, if one has their own datasets, they can perform differential expression analysis and filter the results using the rules described in table 4 of the manuscript. In this case, ensure that the csv files listing the genes are formatted properly (including capitalization and the index (0,1,2…) in the first column):

	inputs	
0	A1BG	
1	A1CF	
2	A2M	

3. Open CRISPR-GEM_Github.ipynb in google colab (Available on Github).
4. Run the ‘Model’ section. Hover your mouse in between the brackets (1) then a play button will appear. Click the play button and let this run. You may need to approve it to access your google drive.



```
#Datasets used for application:
## input_indices=['SRR22560355', 'SRR13018743', 'SRR21009234'] # MSC for chondrogenesis
## input_indices=['SRR2453345', 'SRR2453349', 'SRR2453348'] # iPSC for Treg differentiation
## input_indices=['SRR7107496', 'SRR7107494', 'SRR7107498'] # OA Cartilage

## output_indices=['SRR6337333', 'SRR6337334', 'SRR6337335'] # # Juvenile Cartilage for Chondrogenesis
## output_indices=['SRR7107486', 'SRR7107484', 'SRR7107481'] # Healthy Cart for OA compare
## output_indices=['SRR25253642', 'SRR25253654', 'SRR25253978'] #Tregs for differentiation

prefix='' # Same string you typed into GUI save box,
# Make sure input_list and output_list for that string are moved to GEM folder in google drive
CRISPR_type='' # CRISPRa, CRISPRi, CRISPRko, OR CRISPRki
analyze_methods(prefix, CRISPR_type, input_indices, output_indices)
```

5. Next the model needs to be set up based on the specific application. Use the SRR IDs output by the GUI and chose three random IDs for both the inputs and outputs. Add them in list format for the respective labels (2 & 3).
6. The GUI outputs the input and output genes as *_input_list and *_output_list where * is the designated prefix. The exact same prefix should be pasted between the single quotations after ‘prefix’ (4).
7. Finally, add the CRISPR type to be assessed (5). This should be the same as the one used in the GUI.
8. Run the “Run it…” cell and wait for the model to complete running. The time can range from 2-45 minutes depending on the number of input/output genes and the datasets used.
9. Once the model is finished running, ‘View the results’ can be run to give a ranked list of the top scoring genes (6).

*Don’t forget the datasets you use are very important. Make sure you analyze the datasets using the GUI to make sure you are making meaningful comparisons. If the gene results don’t make sense, new datasets may have to be examined.