Developing Original Cloud-based Bioinformatics Software Applications

Daniel Clarke, Alexander Lachmann, Avi Ma’ayan

Department of Pharmacological Sciences
Mount Sinai Center for Bioinformatics
Icahn School of Medicine at Mount Sinai
New York, NY USA
Drug and gene knockdown followed by genome-wide expression

Transcription factors and histone modifications profiled by ChIP-seq

Drug and knockdown effects on cell viability

KO and mutant genes and their disease phenotypes

Gene expression from patient cohorts with genomics and clinical outcome data

Protein-protein interactions and cell or metabolic pathways

Drugs and toxic chemicals that cause adverse events

Harmonizome

Search for genes or proteins and their functional terms extracted and organized from over a hundred publicly available resources. Learn more.

Example searches
achilles  STAT3  breast cancer
## Functional Associations

STAT3 has 14.374 functional associations with biological entities spanning 8 categories (molecular profile, organism, functional term, phrase or reference, disease, phenotype or trait, chemical, structural feature, cell line, cell type or tissue, gene, protein or microRNA) extracted from 100 datasets.

Click the + buttons to view associations for STAT3 from the datasets below.

If available, associations are ranked by standardized value.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Achilles Cell Line Gene Essentiality Profiles</strong></td>
<td>Cell lines with fitness changed by STAT3 gene knockdown relative to other cell lines from the Achilles Cell Line Gene Essentiality Profiles dataset.</td>
</tr>
<tr>
<td><strong>Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles</strong></td>
<td>Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles</strong></td>
<td>Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray</strong></td>
<td>Tissue samples with high or low expression of STAT3 gene relative to other tissue samples from the Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray dataset.</td>
</tr>
<tr>
<td><strong>Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by RNA-seq</strong></td>
<td>Tissue samples with high or low expression of STAT3 gene relative to other tissue samples from the Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by RNA-seq dataset.</td>
</tr>
<tr>
<td><strong>Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles</strong></td>
<td>Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>Biogps Pathways</strong></td>
<td>Pathways involving STAT3 protein from the Biogps Pathways dataset.</td>
</tr>
<tr>
<td><strong>BioGPS Cell Line Gene Expression Profiles</strong></td>
<td>Cell lines with high or low expression of STAT3 gene relative to other cell lines from the BioGPS Cell Line Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>BioGPS Human Cell Type and Tissue Gene Expression Profiles</strong></td>
<td>Cell types and tissues with high or low expression of STAT3 gene relative to other cell types and tissues from the BioGPS Human Cell Type and Tissue Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>BioGPS Mouse Cell Type and Tissue Gene Expression Profiles</strong></td>
<td>Cell types and tissues with high or low expression of STAT3 gene relative to other cell types and tissues from the BioGPS Mouse Cell Type and Tissue Gene Expression Profiles dataset.</td>
</tr>
<tr>
<td><strong>CCLE Cell Line Gene CNV Profiles</strong></td>
<td>Cell lines with high or low copy number of STAT3 gene relative to other cell lines from the CCLE Cell Line Gene CNV Profiles dataset.</td>
</tr>
<tr>
<td><strong>CCLE Cell Line Gene Expression Profiles</strong></td>
<td>Cell lines with high or low expression of STAT3 gene relative to other cell lines from the CCLE Cell Line Gene Expression Profiles dataset.</td>
</tr>
</tbody>
</table>
Impact of the Ma’ayan Lab Tools

> 41 published bioinformatics tools and databases
> 1.2 million unique users across all tools
> 3,000 unique users per day
> 30,000 unique users per month
> 3,000 papers that cite the tools in 2020
BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface.


BioJupies now supports user accounts! Sign in from the top right corner of the page for access to unlimited private notebooks, RNA-seq datasets and alignment jobs.

https://maayanlab.cloud/biojupies/
The L1000FWD Map - ~17K Signatures, ~5K Drugs

https://maayanlab.cloud/L1000FWD/main

Bioinformatics. 2018 Jun 15;34(12):2150-2152
What are the molecular effects of hydroxychloroquine on human cells?

BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface.

Get Started

https://maayanlab.cloud/biojupies/
Which dataset would you like to analyze?

Use the form below to search 9,145 publicly available datasets published in the Gene Expression Omnibus database and processed by ARCHS4.

hydroxychloroquine

Displaying 1-1 of 1 results | Organism: All | Sort by: Newest | Samples: 6

Hydroxychloroquine inhibits responses to group A streptococcus in peripheral blood mononuclear cells

GSE74235 15 samples Published October 2016

<< 1 > >>
Which analyses would you like to perform?

Use the form below to add or remove data analysis and visualization tools to your notebook. These tools will analyze the selected dataset and embed interactive results in your notebook. Once you have selected the desired tools, click Continue to proceed.

**Exploratory Data Analysis**

These tools assist in visually exploring global patterns within the dataset.

- **PCA**
  Linear dimensionality reduction technique to visualize similarity between samples

- **Clusterogrammer**
  Interactive hierarchical clustering heatmap visualization

- **Library Size Analysis**
  Analysis of readcount distribution for the samples within the dataset

**Differential Expression Analysis**

These tools allow for calculation and exploration of differential gene expression between two groups of samples.

- **Differential Expression Table**
  Differential expression analysis between two groups of samples

- **Volcano Plot**
Which samples would you like to compare?

One or more of the selected tools require generating a gene expression signature. To generate one, you must define two groups of samples whose gene expression you wish to compare by using the form below. Once you have defined the desired groups, click Continue to proceed.

What are the names of the groups?

First, name the groups of samples you wish to compare using the text boxes below. It is recommended to use names descriptive of the groups’ experimental condition; for example, WT, DMSO, Vehicle for the Control group; Metastatic, p53 KO, Dasatinib-6h for the Perturbation group.

Which groups do the samples belong to?

Second, select the samples you wish to assign to the groups by using the dropdown menus below. To improve the statistical validity of the gene expression signature, each group must have at least three samples.
Untreated vs Perturbation Analysis Notebook | BioJupies

Dataset: GSE74235
Signature: Untreated vs Perturbation

Analysis Tools: PCA, Clustergrammer, Library Size Analysis, Differential Expression Table, Enrichr Links, Gene Ontology Enrichment Analysis, Pathway Enrichment Analysis, L1000FWD Query

Tweet  E-mail  Copy Link
Effects of Hydroxychloroquine on Uninfected PBMCs

Untreated vs Perturbation | Gene Ontology Biological Process (2018 version)

Up-regulated in Perturbation
- *cholesterol biosynthetic process (GO:0006695)
- *regulation of alcohol biosynthetic process (GO:1902930)
- *sterol biosynthetic process (GO:0016126)
- *secondary alcohol biosynthetic process (GO:1902653)
- *regulation of cholesterol biosynthetic process (GO:0045540)
- *regulation of cholesterol metabolic process (GO:0090181)
- *regulation of steroid biosynthetic process (GO:0050810)
- *cholesterol metabolic process (GO:0008203)
- *lipid biosynthetic process (GO:0008610)
- *fatty acid derivative biosynthetic process (GO:1901570)
- *sterol metabolic process (GO:0016125)
- *organic hydroxy compound biosynthetic process (GO:1901617)
- *fatty-acyl-CoA biosynthetic process (GO:0046949)
- *acetyl-CoA metabolic process (GO:0006084)
- *acyl-CoA biosynthetic process (GO:0071616)

Down-regulated in Perturbation
- lipid homeostasis (GO:0055088)
- anion homeostasis (GO:0055081)
- cholesterol homeostasis (GO:0042632)
- sterol homeostasis (GO:0055092)
- negative regulation of cholesterol storage (GO:0010887)
- spliceosomal complex assembly (GO:0000245)
- reverse cholesterol transport (GO:0043691)
- phospholipid homeostasis (GO:0055091)
- positive regulation of cell activation (GO:0050867)
- interleukin-1 secretion (GO:0050701)
- regulation of cholesterol transport (GO:0032374)
- mRNA splice site selection (GO:0006376)
- response to lipid (GO:0033993)
- regulation of cholesterol storage (GO:0010885)
- phospholipid efflux (GO:0033700)

https://amp.pharm.mssm.edu/biojupies/notebook/ZxbvpHAn5
The L1000FWD Map - ~17K Signatures, ~5K Drugs

https://maayanlab.cloud/L1000FWD/main

Bioinformatics. 2018 Jun 15;34(12):2150-2152
Projecting the Hydroxychloroquine Signatures onto the L1000FWD Map

Uninfected
Projecting the SARS-CoV-2 Signatures onto the L1000FWD Map

Calu3

A549-ACE2

Drugs Hitting the Same Region in Expression Space

Berbamine

Terfenadine

Loperamide

Amlodipine

Trifluoperazine

RS-504393

Chlorpromazine

A549-ACE2
The L1000FWD Map for the A549 Cell-Line

https://maayanlab.cloud/L1000FWD/graph_page/A549-tSNE_layout.csv

Bioinformatics. 2018 Jun 15;34(12):2150-2152
TenOever Lab (Mount Sinai)

Drugs from Published In-Vitro Screens are Hitting the Same Region in Expression Space

https://appyters.maayanlab.cloud/#/Drugmonizome_Consensus_Terms
Drugmonizome-ML
Views: 377 Starts: 357
Runs: 169 Retrievals: 131
Machine learning pipeline to predict novel drug indications from small molecule attributes
Select

Drugmonizome ETL: DrugRepurposingHub
Views: 47 Starts: 26
Runs: 9 Retrievals: 9
An appyter to process drug-target and drug-mechanism of action associations from the Drug Repurposing Hub.
Select

Drugmonizome Consensus Terms
Views: 45 Starts: 49
Runs: 29 Retrievals: 82
An appyter that queries an input collection of drug sets against libraries in Drugmonizome and returns the top enriched consensus terms
Select

https://appyters.maayanlab.cloud/#/?tags=Drugmonizome
PLK1 Inhibitor as a Potential Drug to Treat Diabetic Kidney Disease

Step 1: Identify DKD studies
Step 2: Extract up/down genes
Step 3: Extract consensus signature
Step 4: Query signatures against the L1000 data
Step 5: Test top reversing drugs in cells and animal models

GEO
Nephroseq

WT+vehicle  Ove26+vehicle  Ove26+B12536

https://appyters.maayanlab.cloud/#/L1000FWD_Consensus_Drugs

Diabetes. 2020 Oct 16; db200580.
An appyter that queries an input collection of drug sets against libraries in Drugmonizome and returns the top enriched consensus terms.
Appyters: turning Jupyter Notebooks into data-driven web apps


Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA
Pencil Worx Design, 345 West 88th Street, New York, NY 10024, USA
Lead contact
Correspondence: avi.maayan@mssm.edu
https://doi.org/10.1016/j.patter.2021.100213
So What is an Appyter and How You Can Create One?

Construct Jupyter Notebook Meta-Report

Compile

CSV

User Submit Data and Parameters

Appyter

Customized & Persistent Report

Developing an Appyter: Start a jupyter-notebook

(venv) my-first-appyter % pip install appyter
...
(venv) my-first-appyter % jupyter notebook .
...
[1 16:46:10.066 NotebookApp] The Jupyter Notebook is running at:

Developing an Appyter: Start the Appyter Server

(venv) my-first-appyter % appyter --profile=biojupies Appyter\ Tutorial.ipynb
...
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

Use Appyter Magic for jinja2 Meta-Programming

Minor changes can turn an existing jupyter notebook into a full-blown web application allowing others to use your data processing pipeline.

```python
#appyter init
from appyter import magic
magic.init(lambda _globals: __())
```

Single Source of Truth: Jupyter Notebook (ipynb)

“Meta” Jupyter Notebook
appyter.ipynb
appyter.json

FAIR Metadata
author
version
license
tags

Version controlled & validated for appyter-catalog

Web Application & REST API

Extract Input Schema from notebook

Render form to get input from user

Render appyter cells with input to produce notebook

Execute notebook in real time

Serve static executed notebook

Dockerize

Multi-Appyter Orchestration

Platform Agnostic Multi-Appyter Orchestration

Ingress (L7 LB)

- S3
- Appyter
- PostgREST
- Catalog
- Appyter Execution
- Appyter Orchestrator
- Postgres

The Appyters Catalog Currently Contains 77 Appyters


https://appyters.maayanlab.cloud/
ARCHS^4 Resource

- Samples can be searched by:
  - Meta data, text annotation from SRA
  - Data driven sample query (highly expressed and low expressed genes)
  - Sample query through functional enrichment
  - Manual selection
Gene Expression Atlas for Tissues and Cell Lines
Prediction of Gene Function using ARCHS4 Data
Prediction of Gene Function using ARCHS4 Data
Mining Massive Gene Expression Repositories

• ARCHS4 was used to align 16 trillion reads ($16 \times 10^{12}$) to human and mouse reference genomes
• > petabyte of data (1 million GB)
• 900,809 samples from 23,739 experiments
Cost challenges

• CPU
  • Hardware purchases are high upfront and require maintenance
  • CPU rentals are high if used continuously

• Memory
  • Aligning reads to a reference genome requires efficient in memory index structures
  • Depending on alignment algorithm this can vary from 4GB (kallisto/De Bruijn Graph) to 16GB (STAR/suffix tree)

• Storage
  • Raw read files need to be stored for processing
  • Required storage varies significantly between samples

• Network bandwidth
  • Fast networking is required to retrieve raw read data
  • Slow bandwidth will increase overall processing time
Hybrid RNA-seq Pipeline Design

- Processing pipelines are dockerized for easy deployment to resource pool
- Control server hosting job commands serves job descriptions upon request of worker nodes
- Job descriptions are JSON objects
- Results are stored at an online location accessible via URL
Hybrid RNA-seq Pipeline Design

- Parallelization through deployment of dockerized workflows
- Instances are precisely chosen for required resources
  - memory
  - bandwidth
  - storage
Cost comparison of RNA-seq resequencing projects

- The average cost of processing an RNA-seq sample is below $0.01
- Minimal resource allocation and low-cost high network bandwidth of cloud resources is a key advantage of the ARCHS4 pipeline
- Choice of efficient alignment algorithm (kallisto) results in low memory footprint

<table>
<thead>
<tr>
<th>RNA-seq resource</th>
<th>ARCHS4</th>
<th>Recount</th>
<th>Toil Recompute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human samples</td>
<td>84,863</td>
<td>61,350</td>
<td>19,931</td>
</tr>
<tr>
<td>Mouse samples</td>
<td>103,083</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total samples</td>
<td>187,946</td>
<td>61,350</td>
<td>19,931</td>
</tr>
<tr>
<td>Cost per sample</td>
<td>&lt; $0.01</td>
<td>$0.73</td>
<td>$1.30</td>
</tr>
</tbody>
</table>
Mount Sinai Center for Bioinformatics

2021 Summer Research Training Program in Biomedical Big Data Science

Research intensive 10-week training program for undergraduate and master’s students

Program Dates: June 7 - August 13, 2021

Students selected for summer session 2021 of our research training program in the Ma’ayan Laboratory at the Icahn School of Medicine at Mount Sinai will conduct faculty-mentored independent research projects in the following areas:

- Data Harmonization
- Machine Learning
- Cloud Computing
- Dynamic Data Visualization

Application Deadline:
February 1, 2021 at 5 PM Eastern Time

Who Should Apply:
Students majoring in Computer Science, Informatics, Mathematics, Statistics, Physics, Engineering, Chemistry/Chemical Sciences or Biological Sciences and have an interest in Biomedical Big Data Science.

Faculty Mentor and Principal Investigator:
Avi Ma’ayan PhD, Professor and Director Mount Sinai Center Bioinformatics Icahn School of Medicine at Mount Sinai New York, New York

Trainee Salary:
$8,000 salary for the 10-week training period

Eligibility:
To be considered for this program, applicants must be:
- U.S. citizen or U.S. permanent resident
- Undergraduate or master’s student in good academic standing

Contact:
Sherry Jenkins, MS
Program Manager
E-mail: sherry.jenkins@mssm.edu

APPLICATION DETAILS: http://labs.icahn.mssm.edu/maayanlab/summer-research-program/
Summary

• We systematically convert publicly available omics datasets into an abstract format centered on genes and drugs.
• The resources we developed have made a big impact on the research community.
• Appyters can enable the rapid development of bioinformatics applications not just for the Mount Sinai Center for Bioinformatics but also for others at Mount Sinai.
• Machine Learning to impute knowledge about gene and drug functions.
• SignatureCommons is a template system to quickly bring up a data portal to serve data and metadata.
Acknowledgements

Ma’ayan Lab
Sherry Jenkins, MS - Project Manager
Daniel Clarke, MS - Data Science Analyst
Alexander Lachmann, PhD - Assistant Professor
Kathleen Jagodnik, PhD - Postdoctoral Fellow
Minji Jeon, PhD - Postdoctoral Fellow
Megan Wojciechowicz, MS - PhD Student
John Erol Evangelista, MS - Bioinformatician
Sherry Xie, BS - Bioinformatician
Eryk Kropiwnicki, MS - Bioinformatician
Maxim Kuleshov, MS - Bioinformatician
Ingrid Shu, BS - Bioinformatician
Allison Bailey, MPH - Associate Researcher