Code-free bioinformatics pipeline for building biological databases used to construct cardiovascular-SNPDB: an online database for SNPs associated with cardiovascular metabolites

Mohammad M. Tarek[1]

[1]Bioinformatics Department, Armed Forces College of Medicine (AFCM), Cairo, Egypt.

E-mail Address: mohammadtareq459@gmail.com

Abstract

Biological databases are of great importance for managing biological research data. Building databases has been a code-based process that requires integrative coding skills of different languages. Herein, we present a code-free pipeline that helps biologists to build their databases with no need for coding skills providing searchable downloadable and editable databases using Google Apps. We provided an example for an online tool including a database of SNPs associated with cardiovascular Metabolites, allowing basic features like browsing, downloading, filtering and printing. We also described a stepwise pipeline for building such an interactive database. Cardiovascular-SNPDB was made available at : https://sites.google.com/view/cvdsnpdb/browse/.

Introduction

Databases are crucial for management of bioinformatics research and tools. Databases manage various biological data types including DNA sequences, Protein Structures and gene expression profiles. Databases usually contain either experimental data, predicted data or both with variety of categories including but not limited to molecules, organisms, molecular pathways or diseases. Databases allow users to browse and download content in different formats. Building bioinformatics databases requires significant programming skills, computer scientists usually use programming languages including PhP, Perl, Python or other languages in a connection with database management system like MySQL and NoSQL. Considering The huge flood of data we are witnessing now in the field of bioinformatics and it is unlimited branches, we need easier and more flexible ways of sharing these data in such way of providing public access to improve communication between researchers and allow better management of datasets. So it would be of great beneficence if we could have a code-free stepwise pipeline that allows biologists of non-programming background to build their online databases and make it accessible to the public without a single line of code.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3189v1 | CC BY 4.0 Open Access | rec: 25 Aug 2017, publ: 25 Aug 2017

Pipeline and Discussions

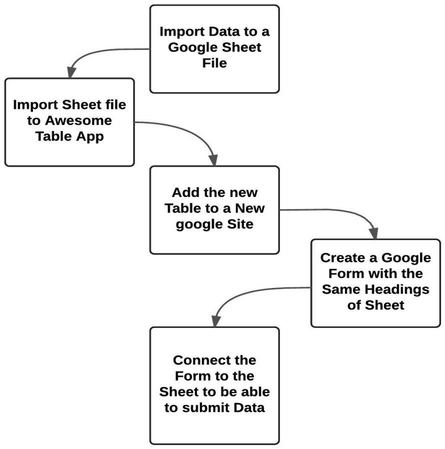
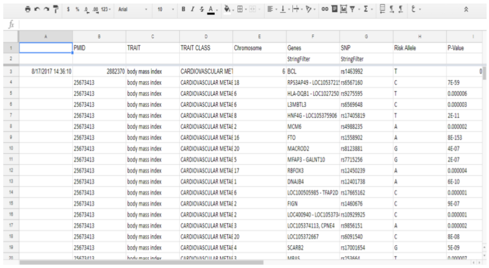

Import Data to a Google Sheet File
Import Sheet file to Awesome Table App
Add the new Table to a New google Site
Create a Google Form with the Same Headings of Sheet
Connect the Form to the Sheet to be able to submit Data

Fig-1 scheme for database building pipeline steps

- **Data input into Google Sheet**

Submitting biological dataset to Google sheets enable authors to easily manage their data on the cloud, so the first step we have used in our pipeline is to import our data we have obtained out from this genetic study[1] into google sheets[2] as shown in figure-2.

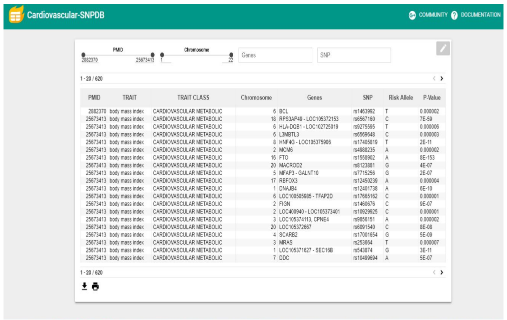


	A	B	C	D	E	F	G	H	I
1		PMID	TRAIT	TRAIT CLASS	Chromosome	Genes	SNP	Risk Allele	P-Value
2						StringFilter	StringFilter		
3	8/17/2017 14:36:10	2882370	body mass index	CARDIOVASCULAR MET	6	BCL	rs1463992	T	0
4		25673413	body mass index	CARDIOVASCULAR METAE	18	RPS3AP49 - LOC10537215	rs6567160	C	7E-59
5		25673413	body mass index	CARDIOVASCULAR METAE	6	HLA-DQB1 - LOC1027250	rs9275595	T	0.000006
6		25673413	body mass index	CARDIOVASCULAR METAE	6	L3MBTL3	rs6569648	C	0.000003
7		25673413	body mass index	CARDIOVASCULAR METAE	8	HNF4G - LOC105375906	rs17405819	T	2E-11
8		25673413	body mass index	CARDIOVASCULAR METAE	2	MCM6	rs4988235	A	0.000002
9		25673413	body mass index	CARDIOVASCULAR METAE	16	FTO	rs1558902	A	8E-153
10		25673413	body mass index	CARDIOVASCULAR METAE	20	MACROD2	rs8123881	G	4E-07
11		25673413	body mass index	CARDIOVASCULAR METAE	5	MFAP3 - GALNT10	rs7715256	G	2E-07
12		25673413	body mass index	CARDIOVASCULAR METAE	17	RBFOX3	rs12450239	A	0.000004
13		25673413	body mass index	CARDIOVASCULAR METAE	1	DNAJB4	rs12401738	A	6E-10
14		25673413	body mass index	CARDIOVASCULAR METAE	6	LOC100505985 - TFAP2D	rs17665162	C	0.000001
15		25673413	body mass index	CARDIOVASCULAR METAE	2	FIGN	rs1460676	C	9E-07
16		25673413	body mass index	CARDIOVASCULAR METAE	2	LOC400940 - LOC105373	rs10929925	C	0.000001
17		25673413	body mass index	CARDIOVASCULAR METAE	3	LOC105374113, CPNE4	rs9856151	A	0.000002
18		25673413	body mass index	CARDIOVASCULAR METAE	20	LOC105372667	rs6091540	C	8E-08
19		25673413	body mass index	CARDIOVASCULAR METAE	4	SCARB2	rs17001654	G	5E-09
20		25673413	body mass index	CARDIOVASCULAR METAE	3	MRAS	rs253664	T	0.000007

Fig-2 Google Sheet containing obtained data

- **Linking Google Sheet into Awesome Table App**

Awesome Table App[3] helps to convert data sheets into searchable applications, so herein we have linked our Google sheet file into Awesome Table App to able to add suitable filters and other options including download and print options into our database App. Figure-2 shows that we have added string filters into both genes and SNPs columns so that users could be able to search these columns' text data easier. Automatically PMID and chromosome columns have been embedded with number range filter. Created Table have been also supplemented with download in *.csv format and print options to supply users with more content accessibility.



PMID 2882370 25673413 Chromosome 1 22 Genes SNP

1 - 20 / 620

PMID	TRAIT	TRAIT CLASS	Chromosome	Genes	SNP	Risk Allele	P-Value
2882370	body mass index	CARDIOVASCULAR METABOLIC	6	BCL	rs1463992	T	0.000002
25673413	body mass index	CARDIOVASCULAR METABOLIC	18	RPS3AP49 - LOC105372153	rs6567160	C	7E-59
25673413	body mass index	CARDIOVASCULAR METABOLIC	6	HLA-DQB1 - LOC102725019	rs9275595	T	0.000006
25673413	body mass index	CARDIOVASCULAR METABOLIC	6	L3MBTL3	rs6569648	C	0.000003
25673413	body mass index	CARDIOVASCULAR METABOLIC	8	HNF4G - LOC105375906	rs17405819	T	2E-11
25673413	body mass index	CARDIOVASCULAR METABOLIC	2	MCM6	rs4988235	A	0.000002
25673413	body mass index	CARDIOVASCULAR METABOLIC	16	FTO	rs1558902	A	8E-153
25673413	body mass index	CARDIOVASCULAR METABOLIC	20	MACROD2	rs8123881	G	4E-07
25673413	body mass index	CARDIOVASCULAR METABOLIC	5	MFAP3 - GALNT10	rs7715256	G	2E-07
25673413	body mass index	CARDIOVASCULAR METABOLIC	17	RBFOX3	rs12450239	A	0.000004
25673413	body mass index	CARDIOVASCULAR METABOLIC	1	DNAJB4	rs12401738	A	6E-10
25673413	body mass index	CARDIOVASCULAR METABOLIC	6	LOC100505985 - TFAP2D	rs17665162	C	0.000001
25673413	body mass index	CARDIOVASCULAR METABOLIC	2	FIGN	rs1460676	C	9E-07
25673413	body mass index	CARDIOVASCULAR METABOLIC	2	LOC400940 - LOC105373401	rs10929925	C	0.000001
25673413	body mass index	CARDIOVASCULAR METABOLIC	3	LOC105374113, CPNE4	rs9856151	A	0.000002
25673413	body mass index	CARDIOVASCULAR METABOLIC	20	LOC105372667	rs6091540	C	8E-08
25673413	body mass index	CARDIOVASCULAR METABOLIC	4	SCARB2	rs17001654	G	5E-09
25673413	body mass index	CARDIOVASCULAR METABOLIC	3	MRAS	rs253664	T	0.000007
25673413	body mass index	CARDIOVASCULAR METABOLIC	1	LOC105371627 - SEC16B	rs543874	G	3E-11
25673413	body mass index	CARDIOVASCULAR METABOLIC	7	DDC	rs10499694	A	5E-07

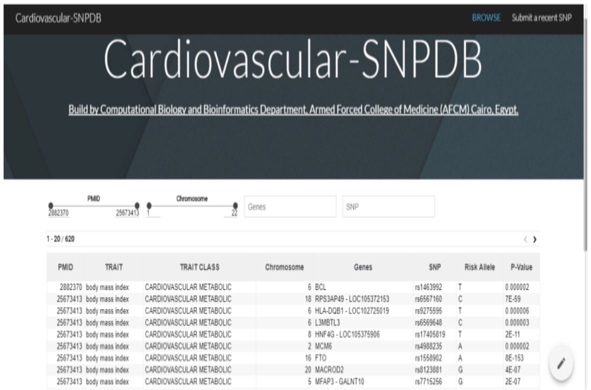
1 - 20 / 620

Fig-3 Awesome Table App Managing the Sheet file

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3189v1 | CC BY 4.0 Open Access | rec: 25 Aug 2017, publ: 25 Aug 2017

- **Building New Google Site Embedded with the Created App**

New Google sites[4] is helpful for easy building of elegant websites that could be supplemented with Google Documents like docs, spreadsheets and forms.



Cardiovascular-SNPDB

Build by Computational Biology and Bioinformatics Department, Armed Forced College of Medicine (AFCM) Cairo, Egypt.

PMID 2882370 25673413 Chromosome 1 22 Genes SNP

1 - 20 / 620

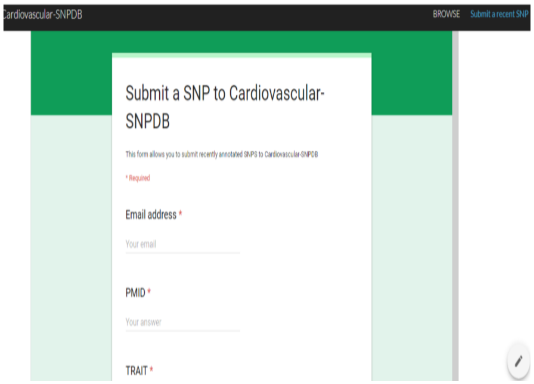
PMID	TRAIT	TRAIT CLASS	Chromosome	Genes	SNP	Risk Allele	P-Value
2882370	body mass index	CARDIOVASCULAR METABOLIC	6	BCL	rs1463992	T	0.000002
25673413	body mass index	CARDIOVASCULAR METABOLIC	18	RPS3AP49 - LOC105372153	rs6567160	C	7E-59
25673413	body mass index	CARDIOVASCULAR METABOLIC	6	HLA-DQB1 - LOC102725019	rs9275595	T	0.000006
25673413	body mass index	CARDIOVASCULAR METABOLIC	6	L3MBTL3	rs6569648	C	0.000003
25673413	body mass index	CARDIOVASCULAR METABOLIC	8	HNF4G - LOC105375906	rs17405819	T	2E-11
25673413	body mass index	CARDIOVASCULAR METABOLIC	2	MCM6	rs4988235	A	0.000002
25673413	body mass index	CARDIOVASCULAR METABOLIC	16	FTO	rs1558902	A	8E-153
25673413	body mass index	CARDIOVASCULAR METABOLIC	20	MACROD2	rs8123881	G	4E-07
25673413	body mass index	CARDIOVASCULAR METABOLIC	5	MFAP3 - GALNT10	rs7715256	G	2E-07

Fig-4 Browse page of the database website created by New Google SItes

then we have supplemented the new website with the public link of the created Awesome Table in a page Named “Browse” so that users could get access to the new database at this link https://sites.google.com/view/cvdsnpdb/.

- **Creating a Google Form for submittable database web Tool**

bioinformatics databases often include data submission in order to keep sustainability of disease related databases to be updates to more recent studies that in this case could discover novel SNPs associated with cardiovascular metabolites. Google form[5] was connected to our Google Sheet that is considered the standard database of the web tool, so that new data submitted in a form containing question boxes that are corresponding to heads of the spreadsheet columns. In this case, each single form answer will be converted to an extra row that will be added to the Database Sheet and should be automatically updated To Awesome Table created App and subsequently to the website of the new database tool. Submission form could be accessed at this link https://sites.google.com/view/cvdsnpdb/submit-a-recent-snp/.



Submit a SNP to Cardiovascular-SNPDB

This form allows you to submit recently annotated SNPS to Cardiovascular-SNPDB

* Required

Email address *

Your email

PMID *

Your answer

TRAIT *

Fig-5 Submission Page on the database website

- **Conclusions**

Bioinformatics is a leading future science that has been transforming biological sciences big data into an easy manageable database driven tools for both biologists and informaticians. This code-free pipeline helps to reduce the coding process that may be an obstacle for biologists who need to make their data more accessible to the scientific community. Integrating Google Apps also helps biologists to manage their created databases on the cloud. This pipeline could be further developed in the future to be more user-friendly and integrate new feature for the best interest of biologists. Our new database could be easily accessed at https://sites.google.com/view/cvdsnpdb/browse/

References:

1. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518(7538):197-206.

2. https://docs.google.com/spreadsheets/.

3. https://awesome-table.com/.

4. https://sites.google.com/.

5. https://docs.google.com/forms/.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3189v1 | CC BY 4.0 Open Access | rec: 25 Aug 2017, publ: 25 Aug 2017