

A dark blue vertical bar on the left side of the slide. A blue arrow points to the right from the bar, containing the date.

4/14/2016

# Opensnp

DSCS 6020 Spring Semester

Several thin, curved lines in dark blue and light gray originate from the bottom left corner and curve upwards and to the right.

Mike Tung

## Contents

Background .....	2
Methods .....	2
Discussion/Conclusion .....	5
References .....	7

## Background

Single Nucleotide Polymorphisms or SNPs are single nucleotide variations that occur at specific regions in a genome. For many years, scientists have conducted studies to characterize SNPs through association studies to determine whether a genetic variant is associated with a disease or trait. (Zhang, et al., 2004)


As mentioned previously by Zhang, et al. SNPs through association studies can determine whether a genetic variation is associated with a disease or trait. One disease of interest is asthma. Asthma is a common long term inflammatory disease of the airways of the lungs and is characterized by variable and recurring symptoms such as reversible airflow obstruction, and bronchospasm (National Heart, Lung, and Blood Institute, 2007). Individuals affected by asthma often cannot exercise heavily and are more reactive to cold air.

Therefore, in my project I present an association study of all the SNPs phenotypes, physical characteristics, associated with asthma and jogging to better understand the severity of the disease with regard to individuals whose life style involves exercise and the life style of individuals who do not exercise stratified by gender.











## Methods

Phenotype characteristics for SNPs were obtained from the Open SNP Project (Open SNP, 2010) using a custom Python script. For each of the user profiles present in the Open SNP Project, the data scraped were outputted in JavaScript Object Notation or JSON format (Figure 1).

openSNP
News
Stats
Genotypes
Phenotypes
SNPs
Search for Everything!
Sign in
FAQ



Bastian Greshake's page

Bastian has uploaded genotyping rawdata.

- Download this set (23andme)
- Download this set (23andme)
- Download this set (23andme-exome-vcf)

On Bastian:

Description

Life Scientist, currently studying ecology and evolution in Frankfurt/Main, Germany and one of the founders of openSNP. Feel free to message me if you encounter bugs.

Homepages

Twitter: <http://www.twitter.com/gedankenstuecke>  
openSNP blog: <http://opensnp.wordpress.com>

Bastian's variations

Characteristic	Variation
white skin	Caucasian
Lactose intolerance	lactose-tolerant
Eye color	blue-green
Hair Type	straight
Height	Tall ( >180cm )
Ability to Tan	Yes
Short-sightedness (Myopia)	low
Beard Color	Blonde
Colour Blindness	False

Figure 1.  
Profile Page OpenSNP User 1 with target data enclosed in red.

Each user profile was scraped individually and processed in to a python dictionary data structure prior to JSON format transformation. (Figure 2) The output JSON file was then manipulated in an R script, makedb.R into a Mongo database of key user and value phenotypes.

```

mike@mike-VirtualBox: ~/github/Open_SNP
[mike-VirtualBox @ Open_SNP]$ ls
asthmaJoggers.tiff  mongoDB.txt      opensnp.json      output      snp_db_data.txt
makedb.R           opensnp.FULL.json opensnp.py         README.md
[mike-VirtualBox @ Open_SNP]$ export PATH=$PATH:./
[mike-VirtualBox @ Open_SNP]$ opensnp.py --help
usage: opensnp.py [-h] [-o]

optional arguments:
  -h, --help            show this help message and exit
  -o, --outfile          outfile name
[mike-VirtualBox @ Open_SNP]$ opensnp.py --outfile opensnp.RUN
processing user id 1
processing user id 2
processing user id 6
processing user id 8
processing user id 9
processing user id 10
processing user id 11
processing user id 13
processing user id 14
processing user id 15

```

Figure 2. Terminal Screen showing sample run of opensnp.py with output JSON name and script usage.

Following the creation of the database, the data was subset with a set of MongoDB queries before visualization of results (Figure 3 & Figure 4).

```

File Edit Code View Plots Session Build Debug Tools Help
Source on Save
24 mongo_data(mongo)
25 mongo_data <- mongo("data")
26 mongo_data$insert(data)
27
28 #check data integrity by export
29 mongo_data$exportToFile("snp_db_data.txt")
30
31 #I'm interested in looking at the distribution of Joggers both male and female with asthma
32 females <- mongo_data$find({"Sex":"female"}, fields = {"Sex":"", "Handedness":"", "Asthma":"", "Jogger":"", "ethnicity":""})
33 males <- mongo_data$find({"Sex":"male"}, fields = {"Sex":"", "Handedness":"", "Asthma":"", "Jogger":"", "ethnicity":""})
34
35 athletes <- rbind(females,males)
36
37 #load up data here to processing related categories together and removing NA's
38 athletes <- na.omit(athletes)
39 athletes$Sex[athletes$Sex == "female"] <- "female"
40 athletes$Jogger[athletes$Jogger == "no"] [athletes$Jogger == "no"] <- "never"
41 athletes$Jogger[athletes$Jogger == "rare"] [athletes$Jogger == "rare"] <- "never"
42 athletes$Jogger[athletes$Jogger == "I work hard and walk alot no need to Jog"] <- "never"
43 athletes$Jogger[athletes$Jogger == "never"] <- "never"
44 athletes$Jogger[athletes$Jogger == "regular"] [athletes$Jogger == "regular"] <- "sometimes"] <- "sometimes"
45
46 athletes$Asthma[athletes$Asthma == "slight"] <- "slight"
47
48 athletes$Asthma[athletes$Asthma == "false"] [athletes$Asthma == "no"] [athletes$Asthma == "no"]
49 athletes$Asthma[athletes$Asthma == "no, don't breathe out much"] <- "no asthma"
50
51 # Run the script
52
Console - github/Open_SNP
> # Make the mongoDB
> mongo_creatdb()
Error: object 'mongo_data' not found
> mongo_data <- mongo("data")
> mongo_data$insert(data)
Complete! Processed total of 4254 rows.
[1] TRUE
>
> #check data integrity by export
> mongo_data$exportToFile("snp_db_data.txt")
Done! Exported a total of 32762 lines.
#
> #I'm interested in looking at the distribution of Joggers both male and female with asthma
> females <- mongo_data$find({"Sex":"female"}, fields = {"Sex":"", "Handedness":"", "Asthma":"", "Jogger":"", "ethnicity":""})
Imported 89 records. Simplifying into dataframe...
> males <- mongo_data$find({"Sex":"male"}, fields = {"Sex":"", "Handedness":"", "Asthma":"", "Jogger":"", "ethnicity":""})
Imported 32762 records. Simplifying into dataframe...

```

Figure 3. Makedb.R Sample Run of MongoDB creation and querying in RStudio IDE

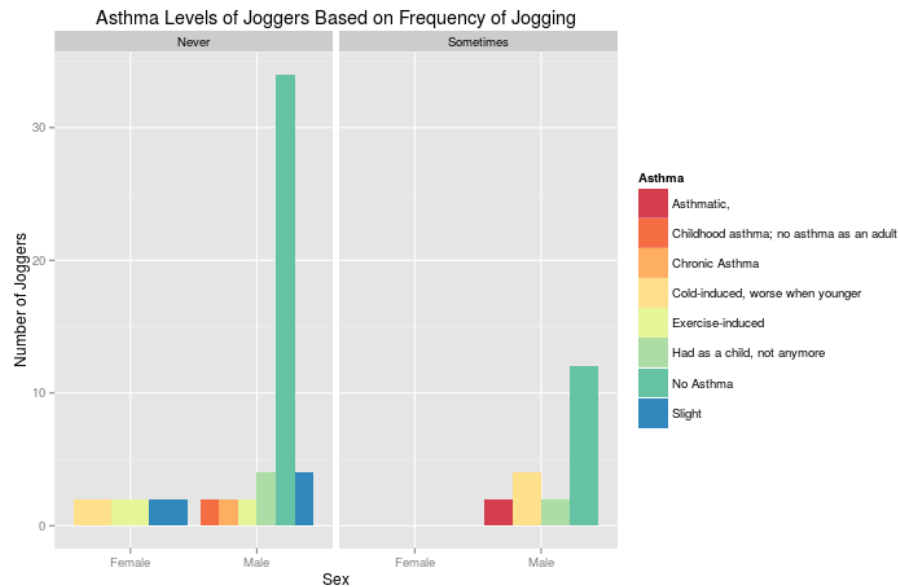


Figure 4.

Faceted Histogram of Female and Male Joggers by Frequency of Jog and Asthma severity represented by color.

## Discussion/Conclusion

One of the initial problems in the search for phenotypic SNPs was building an appropriate web scraper to obtain the data from the open SNP project. Initially after attempting to troubleshoot failed runs in R, it became apparent that R was not the best tool suited to scraping and collecting the enormous amount of data from each of the users in open SNP. The solution was to instead use Python and its wealth of libraries and functions to quickly and effectively scrape the data without timing out the connection to the webpage.

In addition once the data was obtained, the transition from raw data to data appropriate in a Mongo database was easy because the data was already stored in a dictionary data structure. From the data structure the JSON file was easily created and ported to R.

Following the data port, there was the issue of formatting the data and subsetting the data. The problem was remedied by first taking the JSON file and converting it to a data frame and then removing extraneous characters from the column names such as “..”. Next the data frame was converted to a Mongo database using the mongolite library and a quick quality check was conducted with the “mongo export” statement. After verification of data, there was an issue of getting the asthma data, normalizing the data, and getting the exercise data from the database and visualizing the results. The solution was to make two mongo queries to obtain jogging and asthma data by sex and then merge the two data together in to a master data frame for visualization. To normalize the data, the text containing synonyms of asthma descriptions were reformatted to one text, i.e. no, rare, I don’t know to “No”. Some addition pieces of data were also obtained, but were later ignored in the visualization process to improve clarity of data visualization and to remove a layer of complexity in the data analysis.

In conclusion asthma severity does have slight correlation with frequency of exercise. From figure 4, we see that in the male population, asthma is most damaging to individuals who are inactive in their jogs while those of the female population exhibit little to no severe levels of asthma attacks. Despite having obtained the insight from above, much future is required in order to improve the understanding and illuminate the association between asthma and jogging, specifically for the female population. Overall the data obtained became a gateway to future queries involving SNP phenotype characteristics and what associations exist and in addition the scripts, and work flow can be found on GitHub

([https://github.com/seekheart/Open\\_SNP](https://github.com/seekheart/Open_SNP)).

## References

- National Heart, Lung, and Blood Institute. (2007). *Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma*. U.S. Department of Health and Human Sciences.
- Open SNP. (2010). *Open SNP*. Retrieved from Open SNP: <https://opensnp.org/>
- Zhang, K., Qin, Z. S., Liu, J. S., Chen, T., Waterman, M. S., & Sun, F. (2004). Haplotype Block Partitioning and TagSNP Selection Using Genotype Data and Their Applications to Association Studies. *Genome Research*, 908-916.