This week you will work with large-scale genomic datasets to help us better understand incidental findings in diverse populations, especially those that may be obtained upon routine clinical sequencing (e.g. for HCM). In addition to informing clinical interpretation of sequence variants (a huge issue right now), this will give you a chance to build your R/SQL skills, which will be useful throughout the summer and beyond.

The goal is to create an interactive ‘Incidentalome App’ using the R/Shiny framework. You will use the app to grapple with the Incidentalome generally and understand genetic diversity, as well as a lens to understand the recently reported ‘Resilience’ finding (see literature below).  Please create and share a git repo to save/version all code you write (commit regularly!). Below are specific steps to get you started (feel free to skip the R/SQL steps you’ve already completed). After you’ve had a chance to build a basic app by completing the steps below, we’ll map out how to study the connections between the Incidentalome, genetic diversity, and ‘Resilience’.

1. For your Science Communication class, summarize the following three papers (attached):  
   a. "The incidentalome: a threat to genomic medicine" by Kohane et al. JAMA 2006  
   b. “Genetic misdiagnoses and the potential for health disparities” by Manrai et al. NEJM 2016  
   c. “Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases” by Chen et al. Nature Biotechnology 2016
2. Install R [[https://www.r-project.org/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.r-2Dproject.org_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=MCa-z2fJ471I_g9W9b154DRVB7O7fqwqaiDVYZiIajU&e=)] and RStudio [[https://www.rstudio.com/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.rstudio.com_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=xTU-K-nbbxHeLusv92emau9APCaykligYYI_-boA4VY&e=)], and complete a basic tutorial to get comfortable with file input/output, manipulating data frames, and basic statistical operations in R. The following R tutorials are quite good: [https://www.codeschool.com/courses/try-r](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.codeschool.com_courses_try-2Dr&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=x25lvK2i8nlDQFZ-TM0MEKZji89esprRM2JTZ_bGIgE&e=) and [http://www.cyclismo.org/tutorial/R/](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.cyclismo.org_tutorial_R_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=6ZdShz9tgqDZLbTB1Tlx5WRJHg7C-Q4um4EEKR2uU_Y&e=)
3. If you don't already have one, download and install a MySQL/Apache/PHP stack e.g. MAMP ([https://www.mamp.info/en/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.mamp.info_en_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=yf79GJctuFEgxnLIBRwWpSBe0qXND8nydq1EUe1ILmw&e=)) or WAMP ([http://www.wampserver.com/en/](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.wampserver.com_en_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=UD7HHy8QudhWxzV-t_flJVRc2ejhkI_bAGOQpvealsI&e=" \t "_blank)). The following SQL tutorials are great ways to get started with SQL: [https://www.codecademy.com/learn/learn-sql](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.codecademy.com_learn_learn-2Dsql&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=DSMIM3kmiAanOHbwHGXbOxvxVqM4XYPbcab3TVgr38A&e=) and [http://dev.mysql.com/doc/refman/5.7/en/tutorial.html](https://urldefense.proofpoint.com/v2/url?u=http-3A__dev.mysql.com_doc_refman_5.7_en_tutorial.html&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=5sII1sTYD-DE2YXDvB483og9Ae-jAMIhJ1QIkrj5sJ0&e=)
4. Write a single paragraph on the following — What assumptions do Kohane et al. make to create Figure 1 in their JAMA 2006 paper? What parameters control the slope/concavity of the curve in Figure 1?
5. Write an R function that reproduces Figure 1 in Kohane et al. Function inputs should include all parameters identified in 4.
6. Familiarize yourself with R/Shiny ([http://shiny.rstudio.com/tutorial/](https://urldefense.proofpoint.com/v2/url?u=http-3A__shiny.rstudio.com_tutorial_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=LXMyUKK9xLRLk_NOsOpY-Crw9hl35DZReRmA9wy1lt0&e=)).
7. Build an interactive app with R/Shiny to plot Kohane et al.’s Figure 1 as a function of the parameters you identified in 4 above.
8. Download the ExAC allele frequency data ([http://exac.broadinstitute.org/](https://urldefense.proofpoint.com/v2/url?u=http-3A__exac.broadinstitute.org_&d=AwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=KkHuGrW60jf_aUsGxZnJT-YqFR7w_SUTOlIJ3836Xg4&s=WAAcHn1z4oyDHKG7_oyMlFRFuuIUeFBBLdO2RMTvn5U&e=)) for genes MYBPC3 and MYH7 (two key HCM genes)
9. Stage the allele frequency data downloaded in 8 in a MySQL database locally.
10. Write a SQL script to query the top variants in MYBPC3 and MYH7 by the absolute allele frequency difference between the AFR and NFE populations. Plot an empirical CDF of all allele frequency differences using R.
11. Use RMySQL to connect your R/Shiny app to your MySQL db, and create basic UI functionality to allow the user to input an allele frequency threshold, and have your app report the fraction of variants in MYBPC3 and MYH7 separately that are differentiated at that threshold between AFR & NFE.

Downloaded MAMP

shell> mysql

Start MAMP local servers

show databases;

use kohane\_lab;

show tables;

describe MYBPC3;

Open connection to local SQL

/Applications/MAMP/Library/bin/**mysql --host=localhost -uroot -proot**

Created tables MYBPC3 and MYH7 with corresponding column titles.

create table MYBPC3 (

NumKey int, Chrom int, Position int, RSID varchar(30), Reference varchar(50), Alternate varchar(50), Consequence varchar(50), Protein\_Consequence varchar(50), Transcript\_Consequence varchar(50), Filter varchar(10), Annotation varchar(30), Flags varchar(20), Allele\_Count int, Allele\_Number int, Number\_of\_Homozygotes int, Allele\_Frequency dec(20,15),

Allele\_Count\_African int, Allele\_Number\_African int, Homozygote\_Count\_African int, Allele\_Count\_East\_Asian int, Allele\_Number\_East\_Asian int, Homozygote\_Count\_East\_Asian int, Allele\_Count\_European\_Non\_Finnish int, Allele\_Number\_European\_Non\_Finnish int, Homozygote\_Count\_European\_Non\_Finnish int,

Allele\_Count\_Finnish int, Allele\_Number\_Finnish int, Homozygote\_Count\_Finnish int, Allele\_Count\_Latino int, Allele\_Number\_Latino int, Homozygote\_Count\_Latino int, Allele\_Count\_Other int, Allele\_Number\_Other int, Homozygote\_Count\_Other int, Allele\_Count\_South\_Asian int, Allele\_Number\_South\_Asian int, Homozygote\_Count\_South\_Asian int);

Cleaned MYBPC3 and MYH7 csv files and loaded into mysql db (Added NULL where needed, removed column names, converted to tab-delimited).

delete from mybpc3;

load data local infile '/Users/jamesdiao/Documents/Kohane\_Lab/Week\_1/MYBPC3\_new.txt' into table MYBPC3 lines terminated by '\n';

load data local infile '/Users/jamesdiao/Documents/Kohane\_Lab/Week\_1/MYH7\_new.txt' into table MYH7 lines terminated by '\n';

Made query for Position, AFR, NFE, Diff:

select Position,

1.0\*Allele\_Count\_African/Allele\_Number\_African as AFR, 1.0\*Allele\_Count\_European\_Non\_Finnish/ Allele\_Number\_European\_Non\_Finnish as NFE, abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/

Allele\_Number\_European\_Non\_Finnish) as Diff

from mybpc3 order by Diff desc limit 20;

+----------+---------+---------+---------+

| Position | AFR | NFE | Diff |

+----------+---------+---------+---------+

| 47371484 | 0.42666 | 0.73433 | 0.30767 |

| 47350553 | 0.59494 | 0.30920 | 0.28574 |

| 47365199 | 0.09256 | 0.33095 | 0.23839 |

+----------+---------+---------+---------+

Exporting SQL Files:

/Applications/MAMP/Library/bin/mysqldump -uroot -proot kohane\_lab Allele\_Diff > dump.sql

Get Diff

select Position, abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/Allele\_Number\_European\_Non\_Finnish) as Diff from mybpc3 having Diff > 0.01 order by 2 desc;

select count(abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/Allele\_Number\_European\_Non\_Finnish)) from mybpc3 having select(abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/Allele\_Number\_European\_Non\_Finnish)) > 0.01;

select NumKey, count(\*)

from mybpc3 where NumKey in (select NumKey from mybpc3 where abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/Allele\_Number\_European\_Non\_Finnish) > 0.01)

select NumKey, count(\*)

from mybpc3 where NumKey in (select NumKey from mybpc3 where abs(1.0\*Allele\_Count\_African/Allele\_Number\_African - 1.0\*Allele\_Count\_European\_Non\_Finnish/Allele\_Number\_European\_Non\_Finnish) > 0.01)

git status

git add .

git commit -m "added files"

git push origin master

$ git remote add origin <URLFROMGITHUB>

R Markdown

Kaggle

Send HST Schedule

Friday 1 pm