The simulated data set, and how it was inspired by the HCHS/SOL

Tamar Sofer

July 2017

The HCHS/SOL

The Hispanic Community Health Study/Study of Latinos

- A longitudinal, community-based study
- Individuals from four study sites:
  - Chicago, Bronx, Miami, San Diego.
- Hispanics/Latinos were sampled via a two-stage study design
  - First, block units were sampled,
  - Then, households,
  - Finally, all or some of household members.
- So that results from association analyses apply to the general population, be protected from confounding bias due to sampling, and have correct standard errors despite correlations between individuals
  - HCHS/SOL analyses use sampling weights;
  - are adjusted to study center;
  - are fit via mixed models or GEEs.

Hispanics/Latinos and genetic ancestry

- Hispanics/Latinos are admixed, with three ancestral populations: European, Amerindian, and African.
- The proportion of genotypes due to each ancestry differ between people and groups.

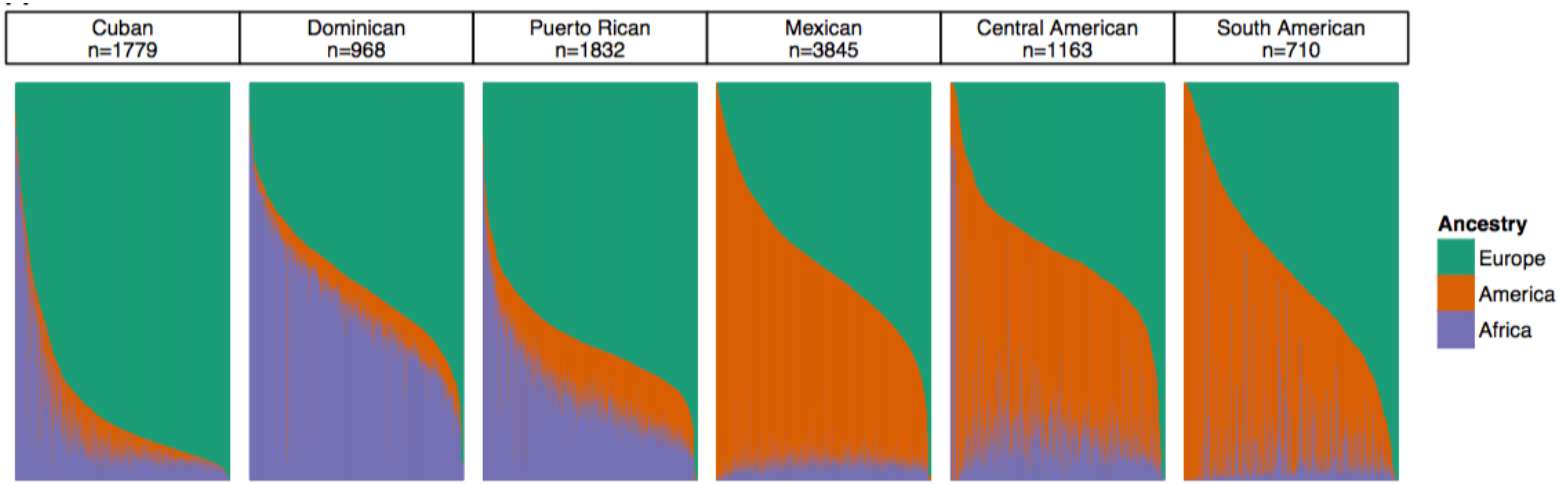
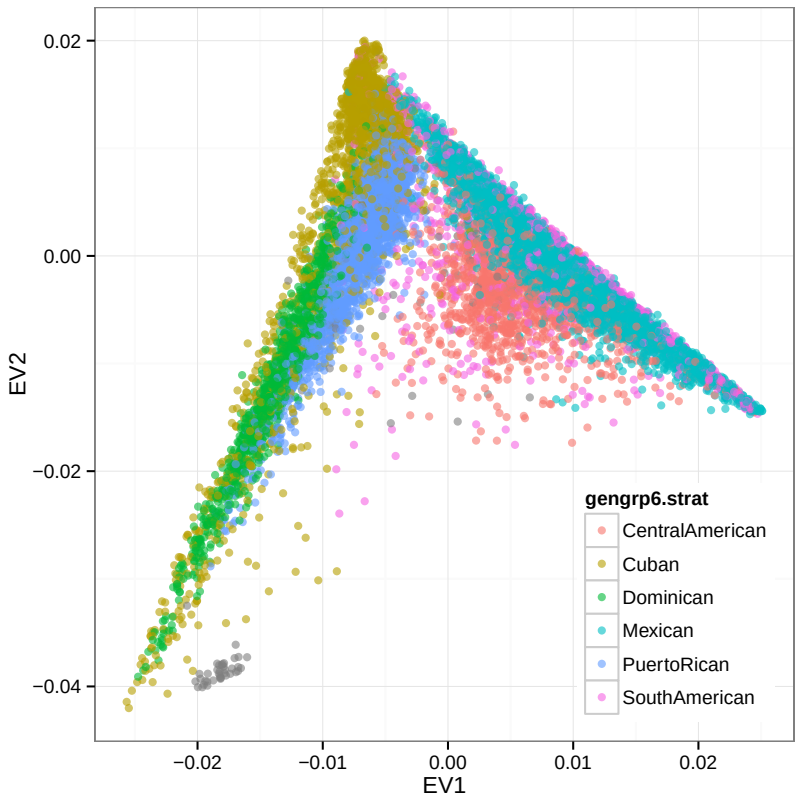


Figure taken from: Conomos, Matthew P., et al. "Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos." The American Journal of Human Genetics 98.1 (2016): 165-184.

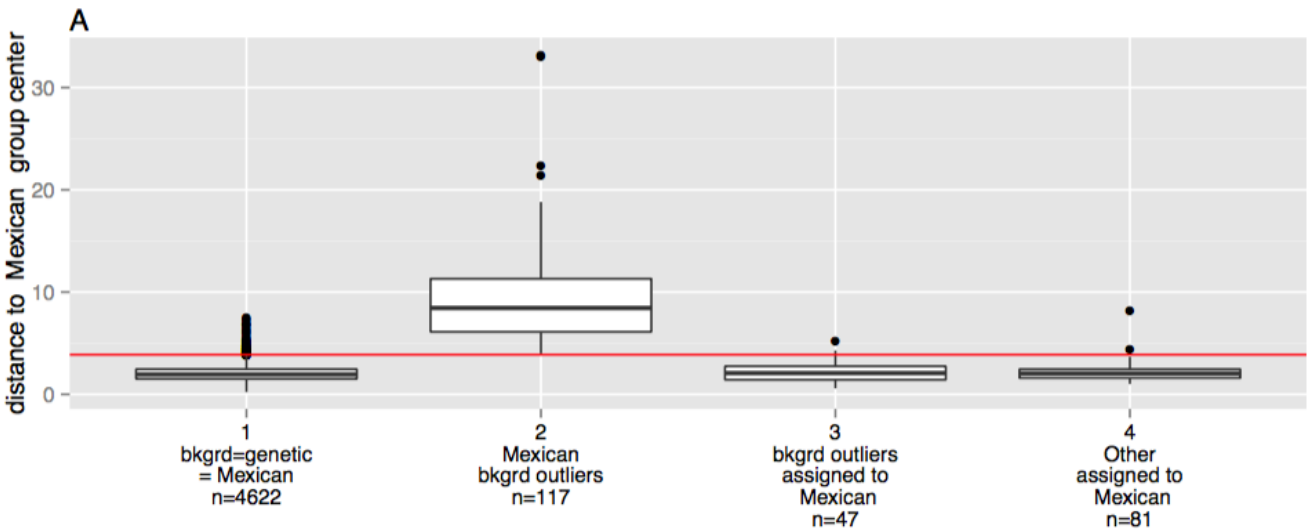
Hispanics/Latinos and genetic ancestry

The diversity of the HCHS/SOL participants and the population structure could also be gleaned from the Principal Components (PCs) figure:



Genetic Analysis Groups

- HCHS/SOL individuals self-identified as Mexican, Central American, South American (Mainland), Cuban, Dominican, or Puerto Rican (Caribbean).
- The HCHS/SOL GAC later defined the Genetic Analysis Groups based on these, and high-dimensional presentation of the genetic data.



- The genetic analysis group is now a factor variable that is used in association analyses in various ways. (How? - later!)

Local ancestry

Due to genetic recombination of chromosomes during Meiosis, genotypes are inherited from parents in intervals.

- Therefore, each chromosome is composed of intervals that were inherited from ancestors.
  - The intervals from more ancient ancestors are smaller.
- Intervals and their ancestries could be inferred using reference panels and an appropriate software.
- For the HCHS/SOL, Browning et al. (2016) performed such inference
  - For each person, we have counts of intervals inherited from each of the parental ancestries.

Browning, Sharon R., et al. "Local Ancestry Inference in a Large US-Based Hispanic/Latino Study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL)." G3: Genes| Genomes| Genetics 6.6 (2016): 1525-1534.

The simulated dataset

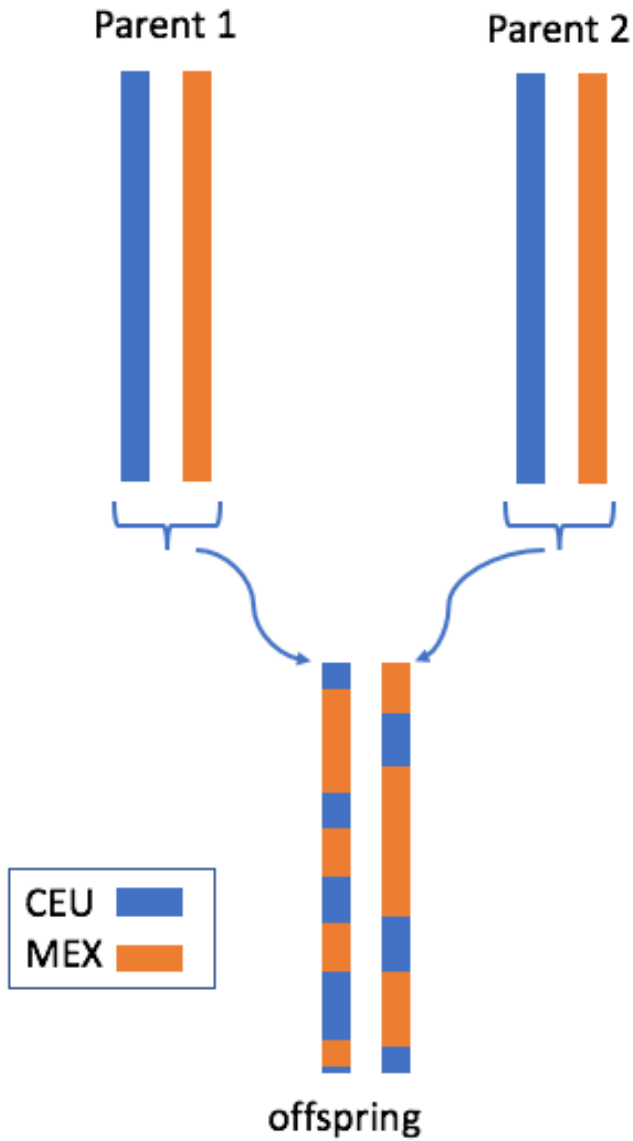
- Due to privacy restrictions, we cannot use the HCHS/SOL dataset.
- So we generated a simulated dataset that have similar, yet simpler, characteristics.
- We will describe the dataset, and then use R packages to study it.

The (simple) simulated dataset

- Using the Hapgen software, simulated genotype data from two populations: CEU and MEX.
  - These are not the same ancestral populations of HCHS/SOL participants.
  - (HCHS/SOL participants have 3 ancestral populations: European, African, Amerindian).
- Assumed that each individual had two parents
  - Of the two chromosome pairs of each parent, one was entirely CEU and one was entirely MEX.
- We randomly assigned intervals inherited from each parent to be those from the first or the second chromosome.
- The probability of CEU ancestry was either 0.8 for the "UW genetic analysis group" or 0.5 for the "UNC genetic analysis group".

The (simple) simulated dataset

- The study individuals have genotypes from both ancestries on each chromosomes.



The simulated data set

- In the remainder of this session, we will look at the simulated data.
- We will get to know a few useful softwares.
- And understand (to some extent) file formats.
- We will not perform any association analysis or testing yet!

Files

- These are the files that we will use.

```
dir <- paste0("/home/postdoc/tsofer/SISG/",
    "Preparing_simulated_data_2")
list.files(dir)
```

```
##  [1] "20170303_prepare_data.R"
##  [2] "20170620_hh_mat.R"
##  [3] "20170705_prepare_gen_data.R"
##  [4] "datasets.zip"
##  [5] "dscvr_diabetes_res.csv"
##  [6] "for_SUGEN"
##  [7] "SISG_genotype.vcf"
##  [8] "SISG_houshold_matrix_2.RData"
##  [9] "SISG_houshold_matrix.RData"
## [10] "SISG_local_ancestry_2.gds"
## [11] "SISG_local_ancestry_snpAnnot.RData"
## [12] "SISG_local_ancestry.gds"
```

The GWASTools R package

If you haven't installed the GWASTools package yet, do so now:

```
source("https://bioconductor.org/biocLite.R")
biocLite("GWASTools")
```

The GWASTools R package

After the package is installed, load it:

```
library("GWASTools", quietly=TRUE)
```

. . . and it may be useful to open the manual

`https://www.bioconductor.org/packages/devel/bioc/manuals/GWASTools/man/GWASTools.pdf`

The GWASTools R package

- GWASTools works with (among the rest) GDS files, which we will use.
- Often, a GDS file will have an "attached" variant annotation file.
- When working with genotype data
  - We first define a genotype reader object [GdsGenotypeReader]
  - Then a genotype data object [GenotypeData]
  - The latter could be associated with the SNP annotation.

Let's see!

Using the GWASTools R package to look at our data

```
gds <- GdsGenotypeReader(file.path(dir,
                                   "SISG_snp_dosages.gds"))
gds
```

```
## File: /home/postdoc/tsofer/SISG/Preparing_simulated_data
## +    [  ]
## |--+ genotype   { Bit2 500x7463, 911.0K }
## |--+ sample.id   { VStr8 500, 2.3K }
## |--+ snp.id   { Int32 7463, 29.2K }
## |--+ snp.chromosome   { Float64 7463, 58.3K }
## \--+ snp.position   { Int32 7463, 29.2K }
```

```
head(getChromosome(gds))
```

```
## [1] 1 1 1 1 1 1
```

```
### the sample IDs of the first 5 individuals
getScanID(gds)[1:5]
```

Using the GWASTools R package to look at our data

```
head(getSnpID(gds))
```

```
## [1] 1 2 3 4 5 6
```

```
head(getPosition(gds))
```

```
## [1] 558390 711153 713682 713754 719811 740098
```

```
### the sample IDs of the first 5 individuals
getScanID(gds)[1:5]
```

```
## [1] "p1" "p2" "p3" "p4" "p5"
```

Using the GWASTools R package to look at our data

Let's connect it to the SNP annotation object via a geotypeData object:

```
snpAnnot <- getobj(file.path(dir,
                             "SISG_snp_dosages_snpAnnot.RData"))
dim(pData(snpAnnot))
```

```
## [1] 7463    9
```

```
head(pData(snpAnnot)[,c(1:5)])
```

```
##         rsID position snpID alleleA alleleB
## 1 rs11497407   558390     1       A       C
## 2 rs12565286   711153     2       G       C
## 3 rs11804171   713682     3       C       T
## 4  rs2977670   713754     4       A       C
## 5  rs2977656   719811     5       A       T
## 6 rs12138618   740098     6       G       A
```

Using the GWASTools R package to look at our data

Let's connect it to the SNP annotation object via a geotypeData object:

```
head(pData(snpAnnot)[,c(6:9)])
```

```
##   chromosome      info type     oevar
## 1          1 1.0000000    2 1.0000000
## 2          1 1.0000000    3 1.0000000
## 3          1 0.8598037    0 0.8992697
## 4          1 0.8786222    0 0.8875563
## 5          1 0.9048905    0 0.9042507
## 6          1 1.0000000    3 1.0000000
```

- "info" and "oevar" are two imputation quality metrics.
- "type" refers to imputation status. type$=0$ is imputed. Otherwise genotyped. (2/3 distinction not important).

Using the GWASTools R package to look at our data

```
varMetadata(snpAnnot)
```

```
##
## rsID
## position
## snpID
## alleleA
## alleleB
## chromosome
## info
## type
## oevar       The oevar imputation quality measure, define
```

Using the GWASTools R package to look at our data

Connecting the genotype reader object it to the snpAnnot object to create a geotypeData object:

```
genoData <- GenotypeData(gds, snpAnnot=snpAnnot)

getAlleleA(genoData)[1:5]
```

```
## [1] "A" "G" "C" "A" "A"
```

```
rsIDs <- getSnpVariable(genoData, "rsID")
rsIDs[1:5]
```

```
## [1] "rs11497407" "rs12565286" "rs11804171" "rs2977670"
```

Using the GWASTools R package to look at our data

```
getGenotypeSelection(genoData, snp = (rsIDs
                                      == "rs2977656"), scan  = 1:10)
```

```
##  p1  p2  p3  p4  p5  p6  p7  p8  p9 p10
##   2   2   2   1   0   1   0   1   2   1
```

Using the GWASTools R package to look at our data

We can also connect the genotype data with sample annotations:

```
scanAnnot <- getobj(file.path(dir,
                              "SISG_phenotypes.RData"))
scanAnnot
```

```
## An object of class 'ScanAnnotationDataFrame'
##   scans: 1 2 ... 500 (500 total)
##   varLabels: scanID EV1 ... group (8 total)
##   varMetadata: labelDescription
```

```
genoData <- GenotypeData(gds,
                snpAnnot=snpAnnot, scanAnnot = scanAnnot)
varLabels(scanAnnot)[1:4]
```

```
## [1] "scanID" "EV1"    "EV2"    "sex"
```

Now we can calculate allele frequencies:

- Only now, because we need to have sex annotation for that!
- . . . the sex column in scanAnnot must be called “sex“. Males have to be denoted by M and females by F.

```
varLabels(scanAnnot)[5:length(varLabels(scanAnnot))]
```

```
## [1] "age"     "trait"   "disease" "group"
```

```
Afreqs <- alleleFrequency(genoData)
```

```
## reading scan 100 of 500

## reading scan 200 of 500

## reading scan 300 of 500

## reading scan 400 of 500

## reading scan 500 of 500
```

Using the GWASTools R package to look at our data

```
head(Afreqs)
```

```
##              M             F   all n.M n.F   n   MAF
## 1 0.00000000 0.000000000 0.000 228 272 500 0.000
## 2 0.00000000 0.001838235 0.001 228 272 500 0.001
## 3 0.01315789 0.003676471 0.008 228 272 500 0.008
## 4 0.00000000 0.000000000 0.000 228 272 500 0.000
## 5 0.80482456 0.779411765 0.791 228 272 500 0.209
## 6 0.84429825 0.810661765 0.826 228 272 500 0.174
```

Using the GWASTools R package to look at our data

```
## close the GDS file:
require(gdsfmt)
showfile.gds(close = TRUE)
```

```
##
## 1 /home/postdoc/tsofer/SISG/Preparing_simulated_data_2/S
##   ReadOnly  State
## 1     TRUE closed
```

or we can also use close(gds).

Other aspects of the data

- Recall that study individuals were sampled from block units, and households.
- These induce correlations between traits of certain individuals
  - E.g. people who live in the same house may eat similar food (similar environment).
- Individuals are also genetically related.
- Similar to the HCHS/SOL, our simulated data set have household and genetic relatedness matrices.
  - The matrices were constructed based on the real data (by sampling from the real matrices).

Other aspects of the data

```
HH.mat <- getobj(file.path(dir,
    "SISG_houshold_matrix.RData"))
kin.mat <-  getobj(file.path(dir,
    "SISG_relatedness_matrix.RData"))
kin.mat[1:5,1:5]
```

```
##                p1           p2           p3           p4
## p1  0.994917413 0.0123160048 -0.0031369458  0.002997052
## p2  0.012316005 0.9962207088  0.0001299051  0.007458538
## p3 -0.003136946 0.0001299051  0.9919895296 -0.002012665
## p4  0.002997053 0.0074585380 -0.0020126658  0.999019918
## p5 -0.004063531 0.0019388646 -0.0037603344  0.000676182
```

Other aspects of the data

- The household matrix have 1 in the i, j entry, if the i, j individuals live in the same household.

```
HH.mat[1:5,1:5]
```

```
##     p1 p2 p3 p4 p5
## p1   1  0  0  0  0
## p2   0  1  0  0  0
## p3   0  0  1  0  0
## p4   0  0  0  1  0
## p5   0  0  0  0  1
```

```
sum(rowSums(HH.mat) > 1)
```

```
## [1] 19
```

Other aspects of the data

- According to this household matrix, only 19 individuals live in the same house as other people in the study.
- There are negative kinship values, and diagonal values are not exactly 1. This is okay.

Exercises

Use the GWASTools manual, your R knowledge, and the commands we learned to perform the following tasks and answer the questions:

1. Compare the variance of the trait "trait" between the UW and the UNC groups.
2. Plot a graph comparing the effect allele frequencies between the groups.
3. What is the genomic position of the SNP with the largest EAF difference between the UW and the UNC groups?
4. What is the proportion of diseased individuals in males and females? and in the UW and UNC groups?
5. Extract the genotypes of rs12033927 and rs17390062. What is the LD between them in the combined sample? in the UW group? in the UNC group?