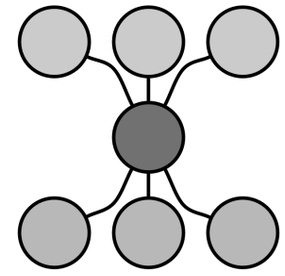


Hands-on tutorial to get you started to do ***X-wide Association Studies (XWASs) with survey data***

Chirag J Patel
(with Nam Pho, Jake Chung, and Arjun Manrai)

ISEE pre-conference tutorial, part 2
Ottawa, Canada
8/26/18

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org

Resources Index
(for today's session)

http://bit.ly/xwas_with_nhanes

Please let us know if you are using the resources
(or provide feedback)!

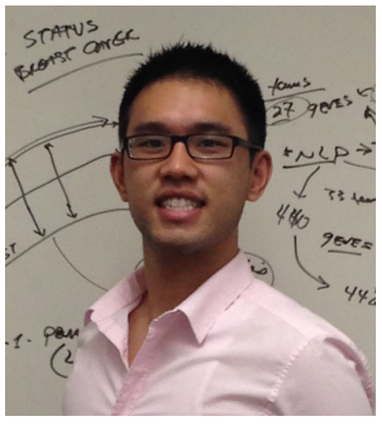
Chirag

@chiragjp

Jake

@jakemkc

Nam

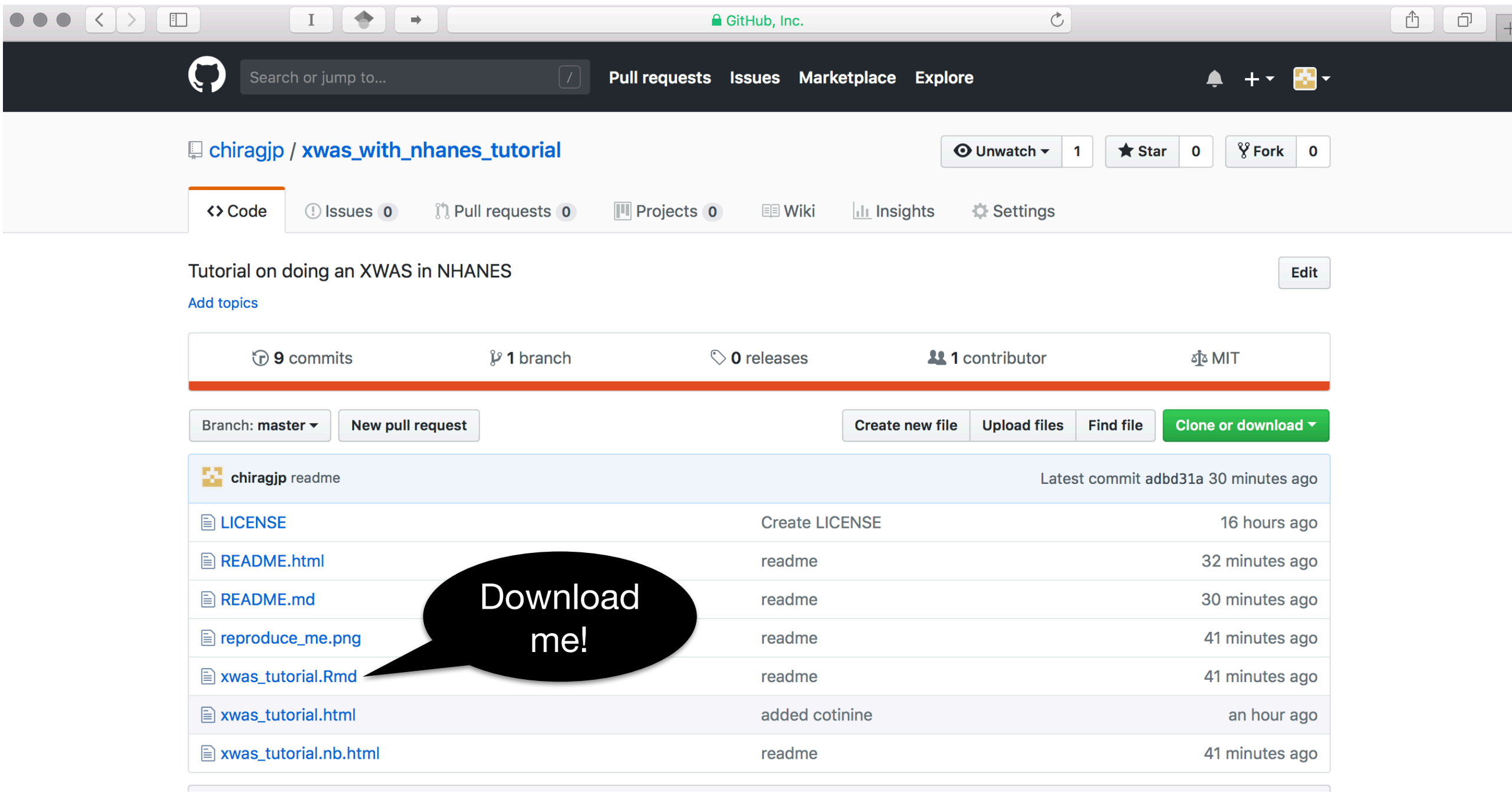


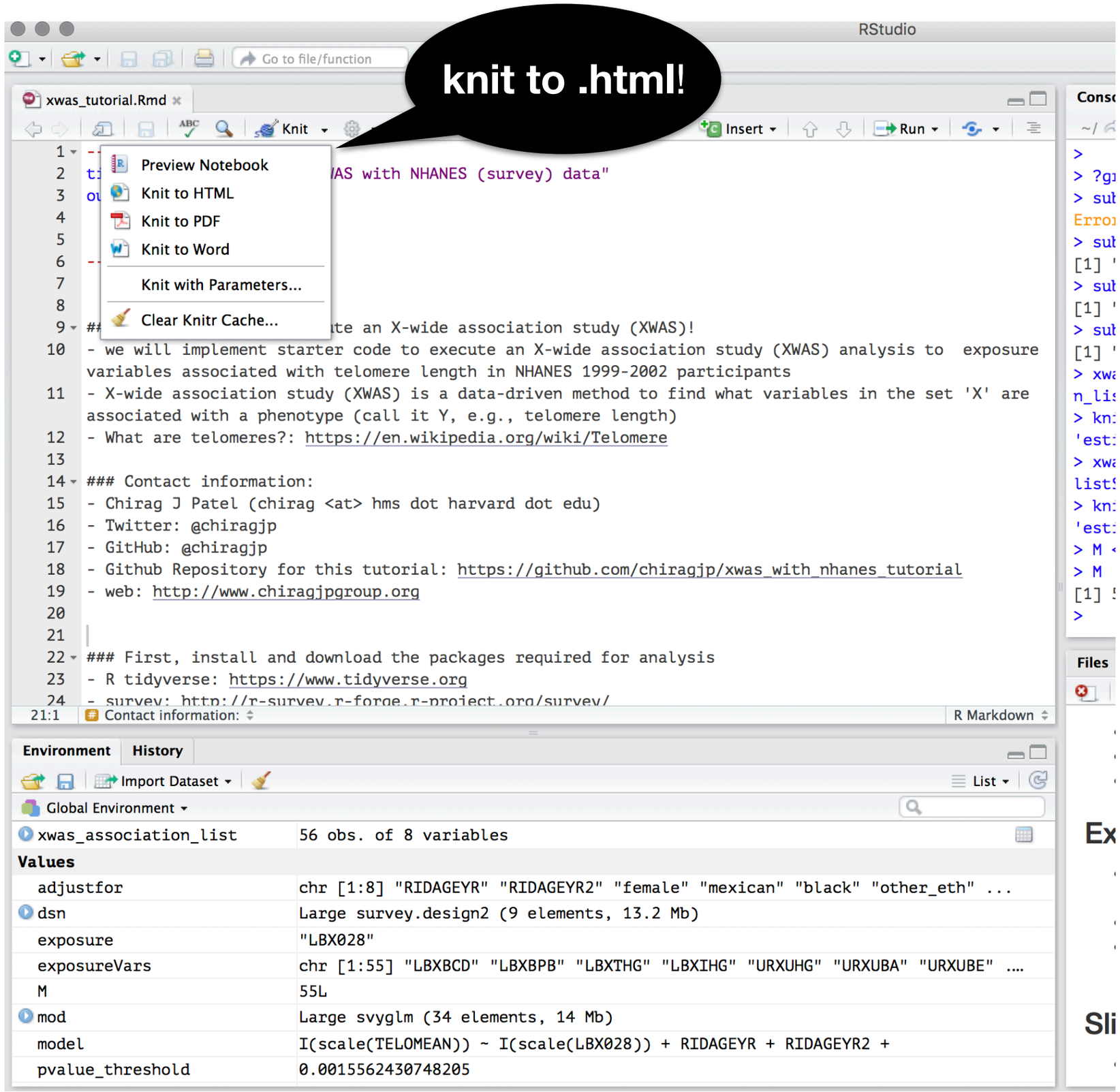
@nampho2

Software prerequisites

- R (version >= 3.3)
- RStudio (version >= 1.01)
- Packages:
  - `survey, broom, tidyverse, knitr`

http://bit.ly/xwas_with_nhanes




RStudio
knit to .html!
Go to file/function
xwas_tutorial.Rmd
Knit
Insert
Run
Preview Notebook
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Clear Knitr Cache...
1 --
2 ti AS with NHANES (survey) data"
3 ou
4
5
6 --
7
8
9 ## te an X-wide association study (XWAS)!
10 - we will implement starter code to execute an X-wide association study (XWAS) analysis to exposure variables associated with telomere length in NHANES 1999-2002 participants
11 - X-wide association study (XWAS) is a data-driven method to find what variables in the set 'X' are associated with a phenotype (call it Y, e.g., telomere length)
12 - What are telomeres?: https://en.wikipedia.org/wiki/Telomere
13
14 ### Contact information:
15 - Chirag J Patel (chirag <at> hms dot harvard dot edu)
16 - Twitter: @chiragjp
17 - GitHub: @chiragjp
18 - Github Repository for this tutorial: https://github.com/chiragjp/xwas_with_nhanes_tutorial
19 - web: http://www.chiragjpgroup.org
20
21
22 ### First, install and download the packages required for analysis
23 - R tidyverse: https://www.tidyverse.org
24 - survey: http://r-survey.r-forge.r-project.org/survey/
21:1
Contact information:
R Markdown
Environment
History
Import Dataset
List
Global Environment
xwas_association_list
56 obs. of 8 variables
Values
adjustfor
chr [1:8] "RIDAGEYR" "RIDAGEYR2" "female" "mexican" "black" "other_eth" ...
dsn
Large survey.design2 (9 elements, 13.2 Mb)
exposure
"LBX028"
exposureVars
chr [1:55] "LBXBCD" "LBXBPB" "LBXTHG" "LBXIHG" "URXUHG" "URXUBA" "URXUBE" ....
M
55L
mod
Large svyglm (34 elements, 14 Mb)
model
I(scale(TELOMEAN)) ~ I(scale(LBX028)) + RIDAGEYR + RIDAGEYR2 +
pvalue_threshold
0.0015562430748205
Files

Real quick:

What is the *exposome*? What is the *phenome*?

exposome

<u>internal</u>

lead (serum)

nutrients (serum)

infection (urine)

metabolome

<u>external</u>

geography

air pollution

income

phenome

<u>function</u>

expression

telomere length

metabolome

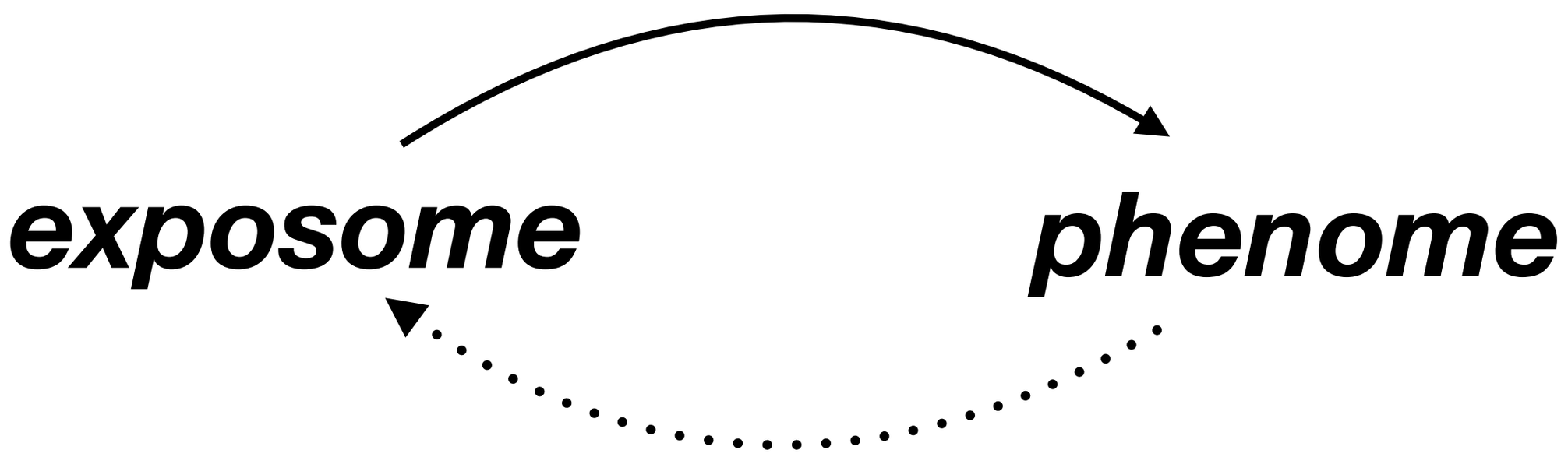
<u>diseases</u>

diabetes

cancer

heart disease

X-wide Association Investigations: Correlating the exposome with the phenome



Recall: ’pseudo-code’ for implementation of an *XWAS: how would you do it?*

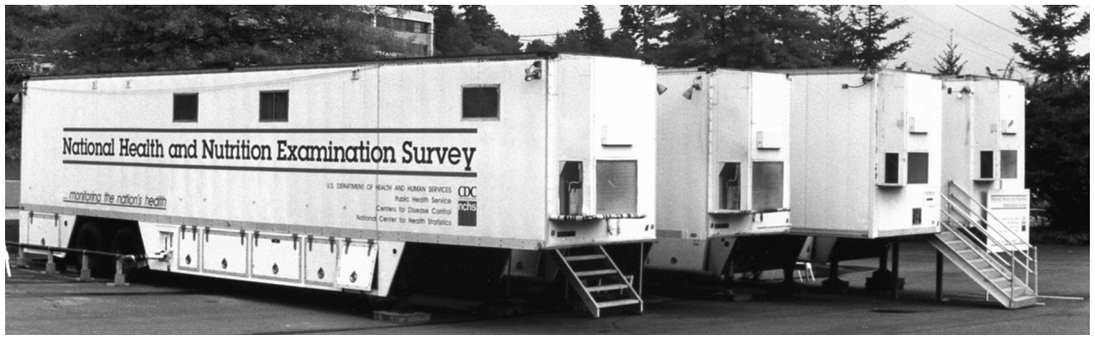
```
y = [blood pressure values for cohort]

association_list = empty_list()

for each x in list of exposures:
   association_test=f(x,y)
   append(association_list, association_test)

multiplicity_correct(association_list)

volcano_plot(y, x, association_list)
```

What is stored in *y*?
What is stored in *association_list*?
What can *f* be?
What does *append* do?
What is the reason for *multiplicity_correct?*

Manrai et al ARPH 2017

Gold standard for ***breadth*** of human exposure information: National Health and Nutrition Examination Survey[1]

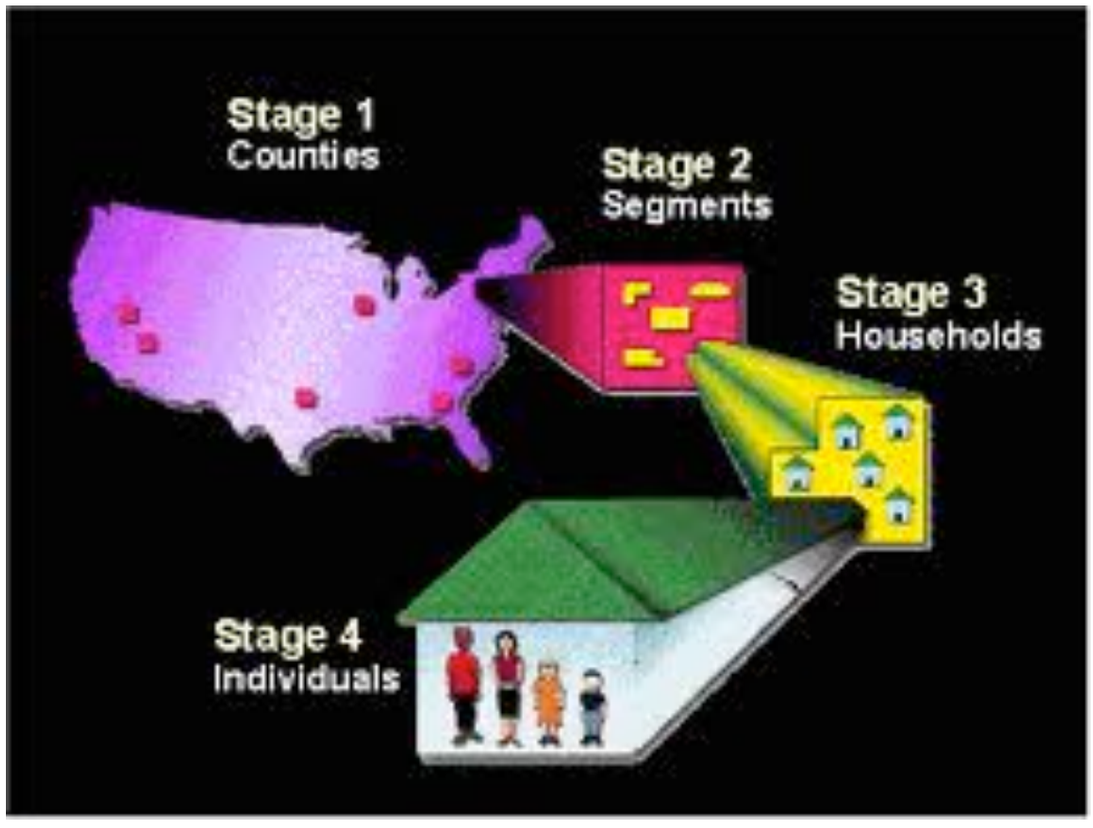
since the 1960s
now biannual: 1999 onwards
10,000 participants per survey



>250 exposures (serum + urine)
GWAS chip

>100s quantitative clinical traits (e.g., serum glucose, lipids, body mass index)

Death index linkage (cause of death)



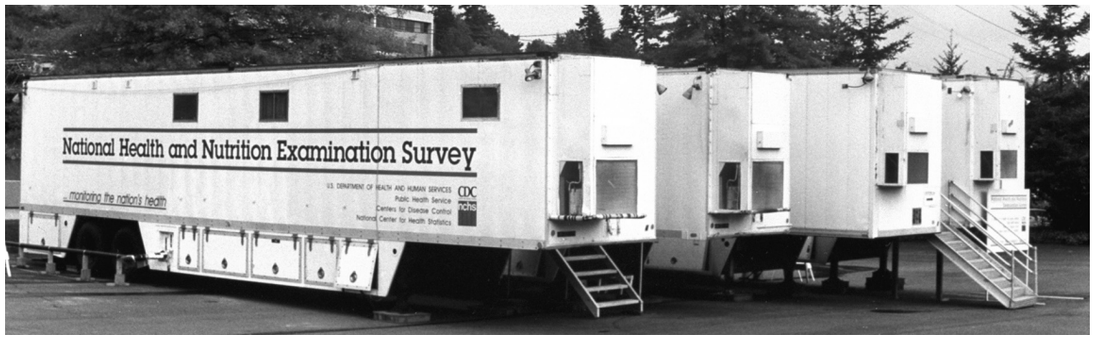
Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

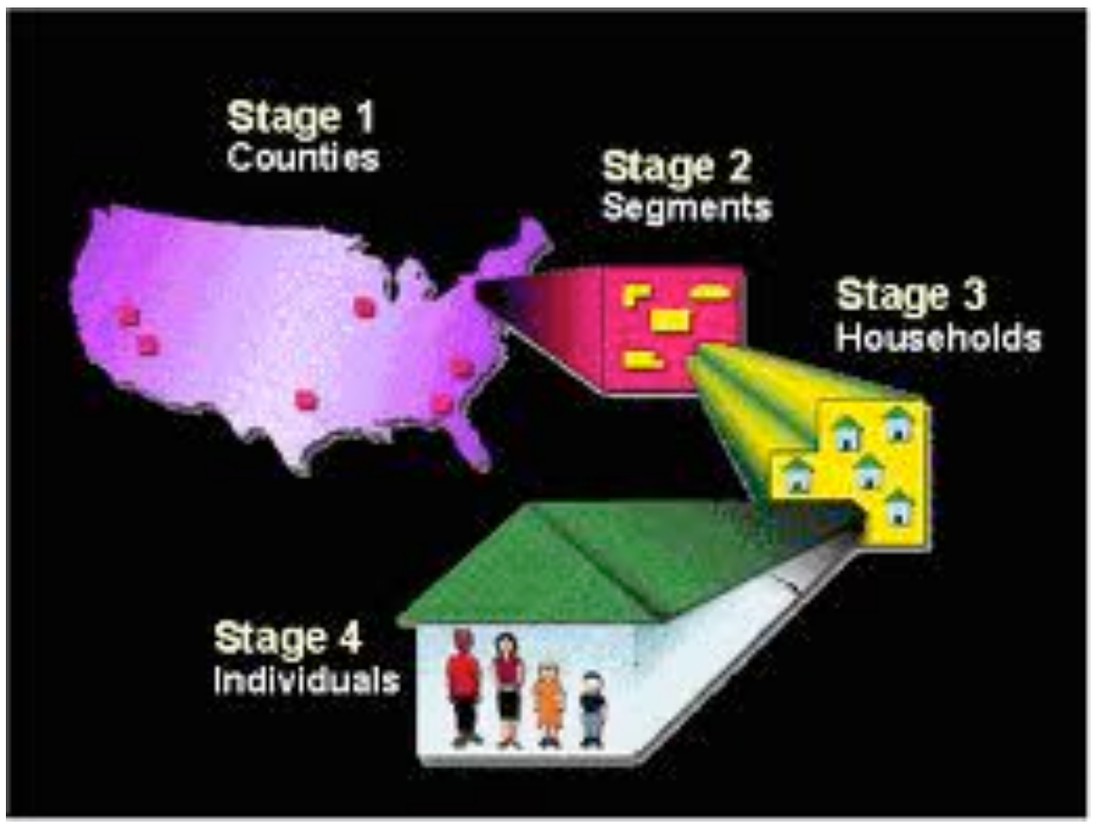
[1] http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***US health and phenotypic monitoring!***

- disease
  - heart disease
  - obesity
  - diabetes

- mortality



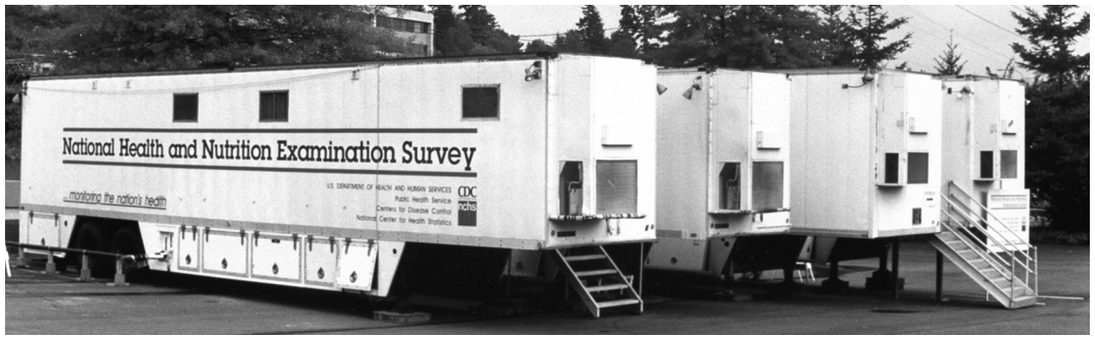


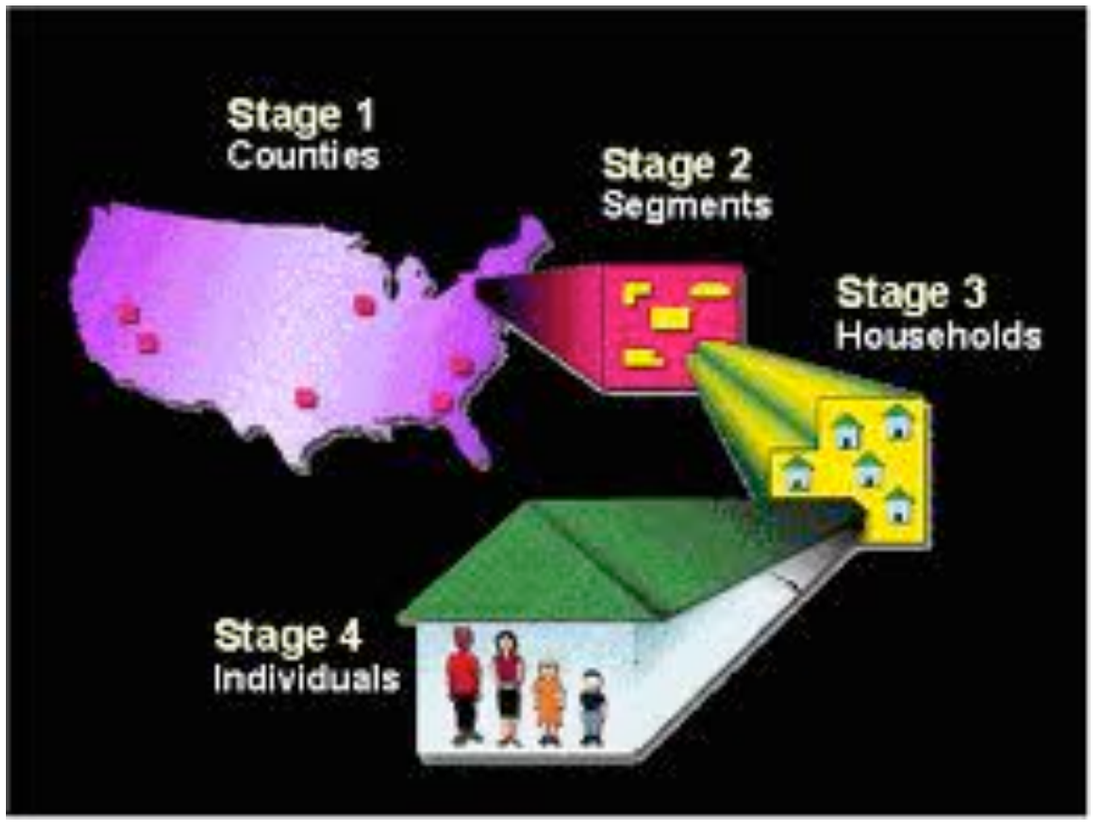
Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

l http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***US health and phenotypic monitoring!***

- risk phenotypes
  - cholesterol
  - BMI
  - glucose
  - lung function
  - kidney function
  - bone mineral density





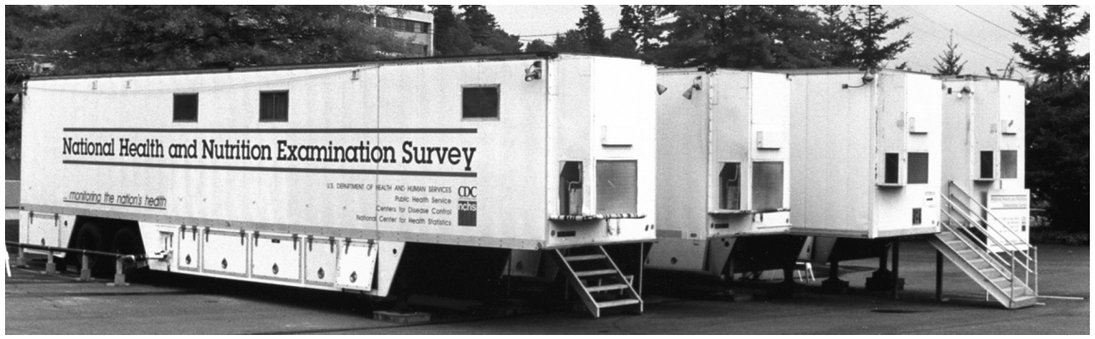
Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

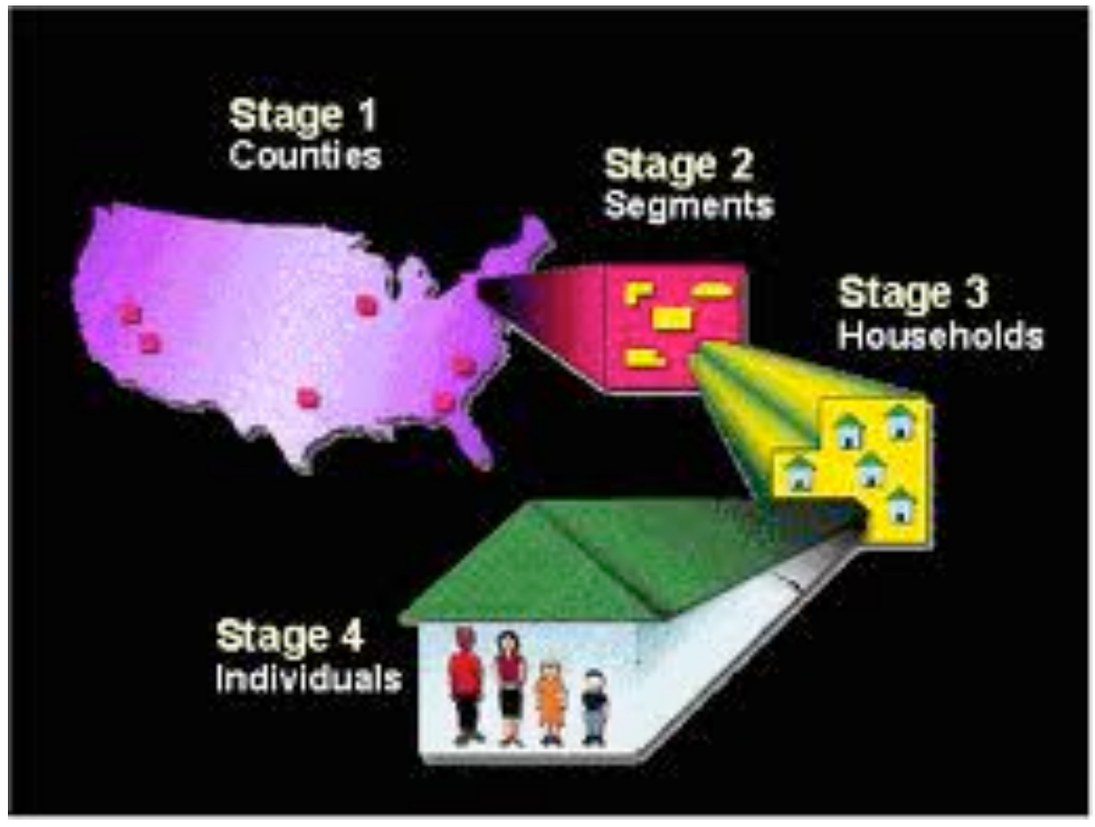
I http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***US health and phenotypic, and <u>exposome</u> monitoring!***

- behavior
  - diet
  - smoking
  - physical activity

- biomarkers of exposures

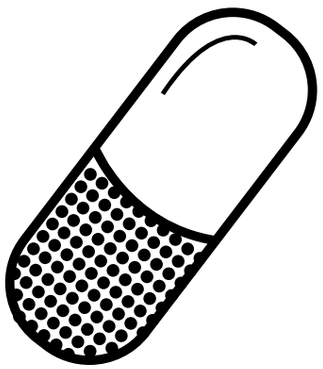




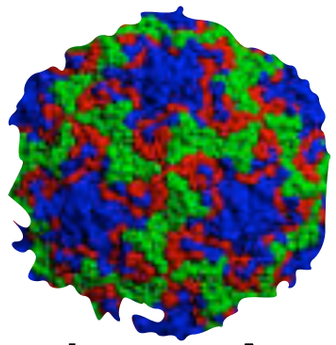
l http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***breadth*** of exposure & behavior data:
National Health and Nutrition Examination Survey

Nutrients and Vitamins
vitamin D, carotenes



Drugs
statins; aspirin



Infectious Agents
hepatitis, HIV, Staph. aureus

Plastics and consumables
phthalates, bisphenol A

Pesticides and pollutants
atrazine; cadmium; hydrocarbons

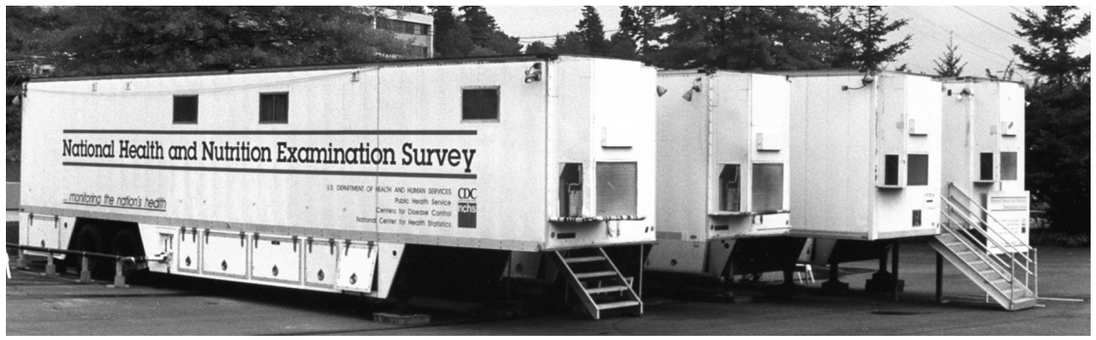
Physical Activity
e.g., steps

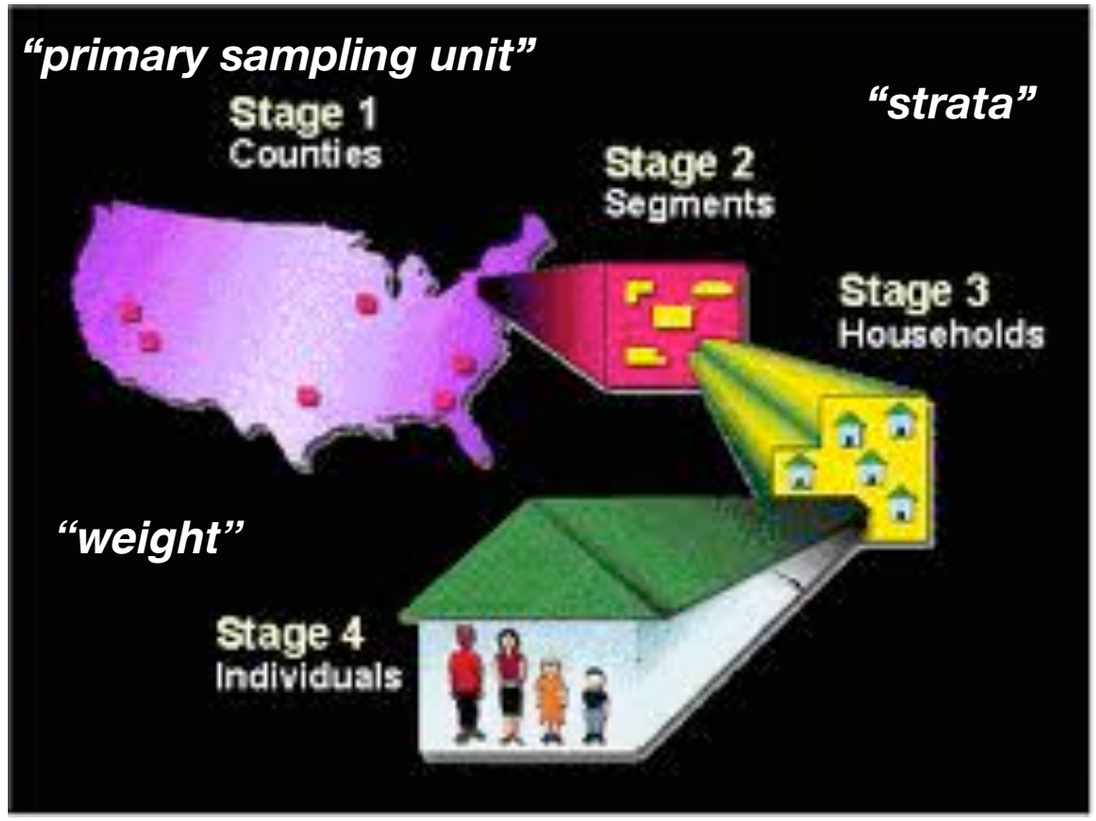
NHANES avoids sampling bias!
Representative of the non-institutionalized population of the United States

15 sampling units per year
- Counties
- Communities
- Households
- Individual

Over-"weights":
- Seniors
- Children
- Ethnic minorities

"***Stratified design***"

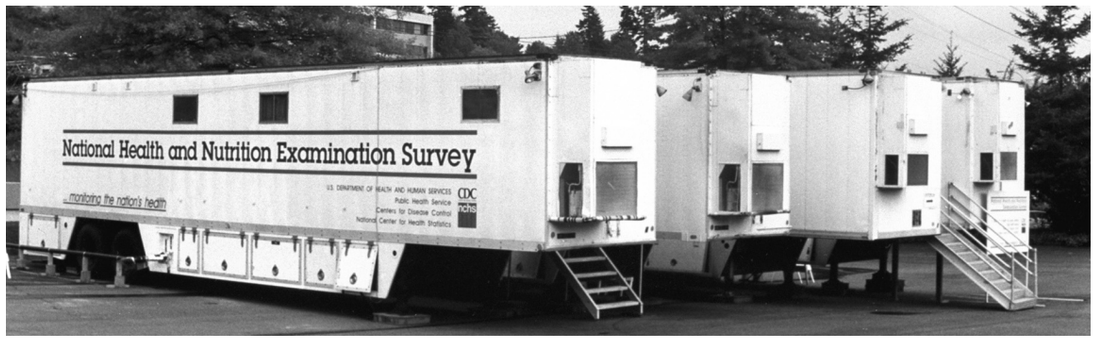


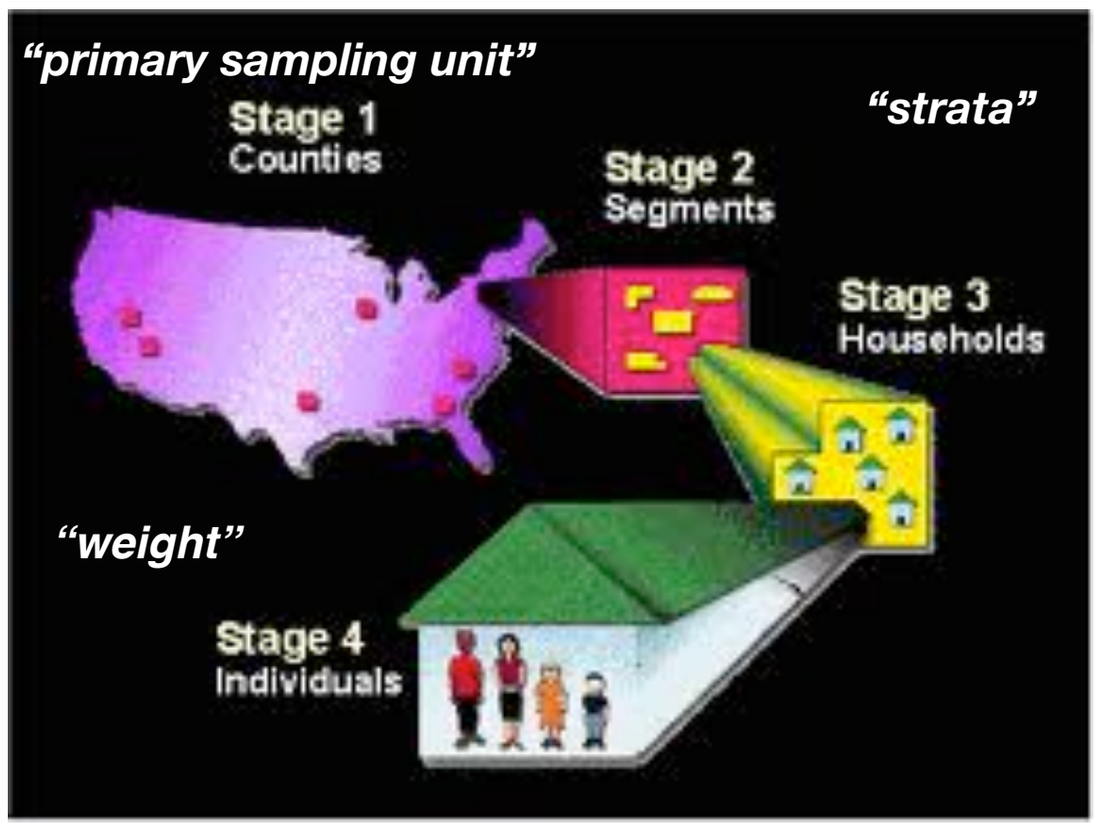


NHANES avoids sampling bias by executing a ***stratified sampling design***

What does this mean?

- Individuals from within each strata are more alike
- (e.g., results from within each strata are correlated)
- why does this matter?



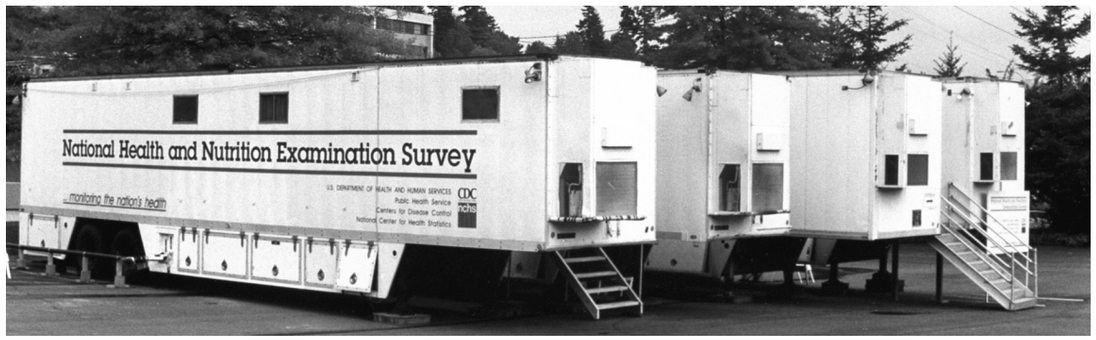


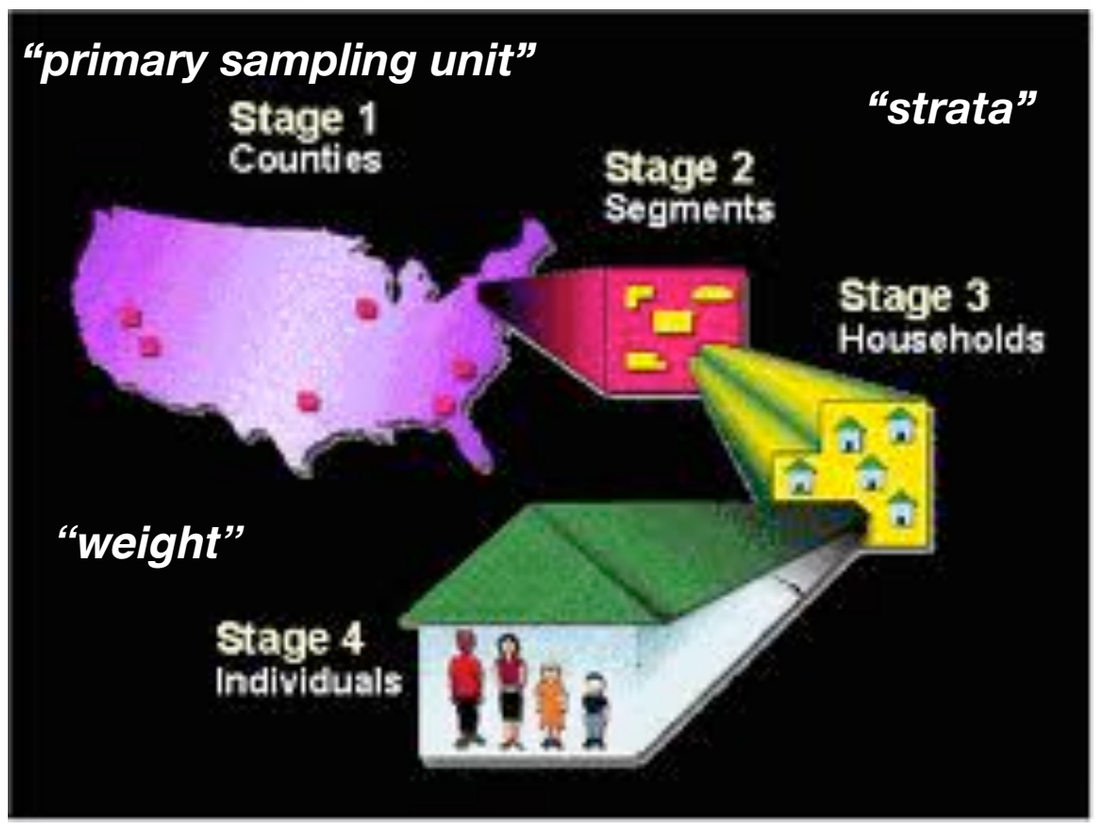
Recall: assumptions of regression and correlated data.

NHANES avoids sampling bias by executing a ***stratified sampling design***

Why does this matter?

- Analytic methods (e.g., regression, t-tests) need to be modified
- Take account of correlation within strata and the probability of being selected



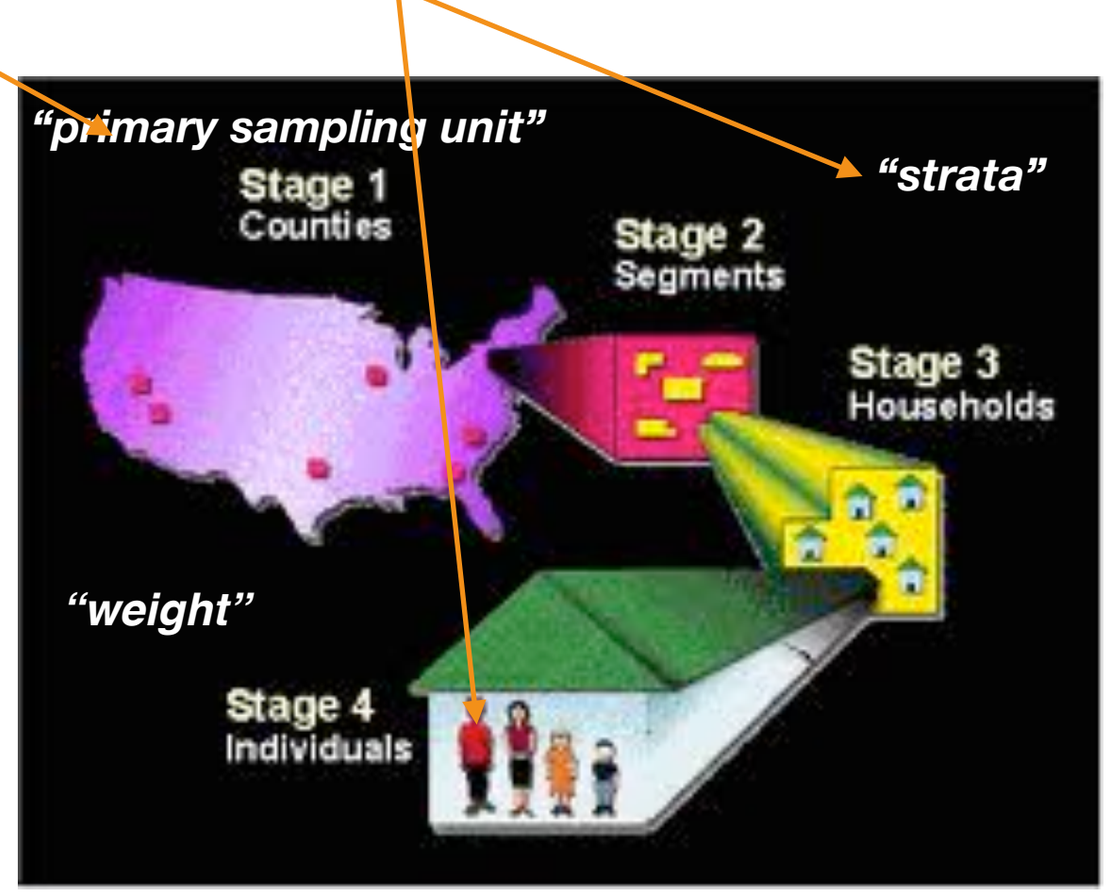


NHANES requires special analytic techniques that account for the stratified design: *survey* package in R

surveyData <- svydesign(ids=~SDMVPSU, strata=~SDMVSTRA, weights=~WTMEC2YR, nest=T, data=nhData.train)

then use the “svy” functions:

- svyglm
- svyttest
- svymean
- svyvar



Knowing how to use the survey ***package is useful!***

Many big surveys use stratified designs.

AMERICAN COMMUNITY SURVEY

U.S. CENSUS BUREAU

The DHS Program

Demographic and Health Surveys

International Journal of Epidemiology, 2016, 1–13
doi: 10.1093/ije/dyw043
Original article

Original article

Systematic correlation of environmental exposure and physiological and self-reported behaviour factors with leukocyte telomere length

Chirag J. Patel,* Arjun K. Manrai, Erik Corona, and Isaac S. Kohane

Department of Biomedical Informatics, Harvard Medical School, Boston. MA, USA

*Corresponding author. Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA. E-mail: chirag_patel@hms.harvard.edu

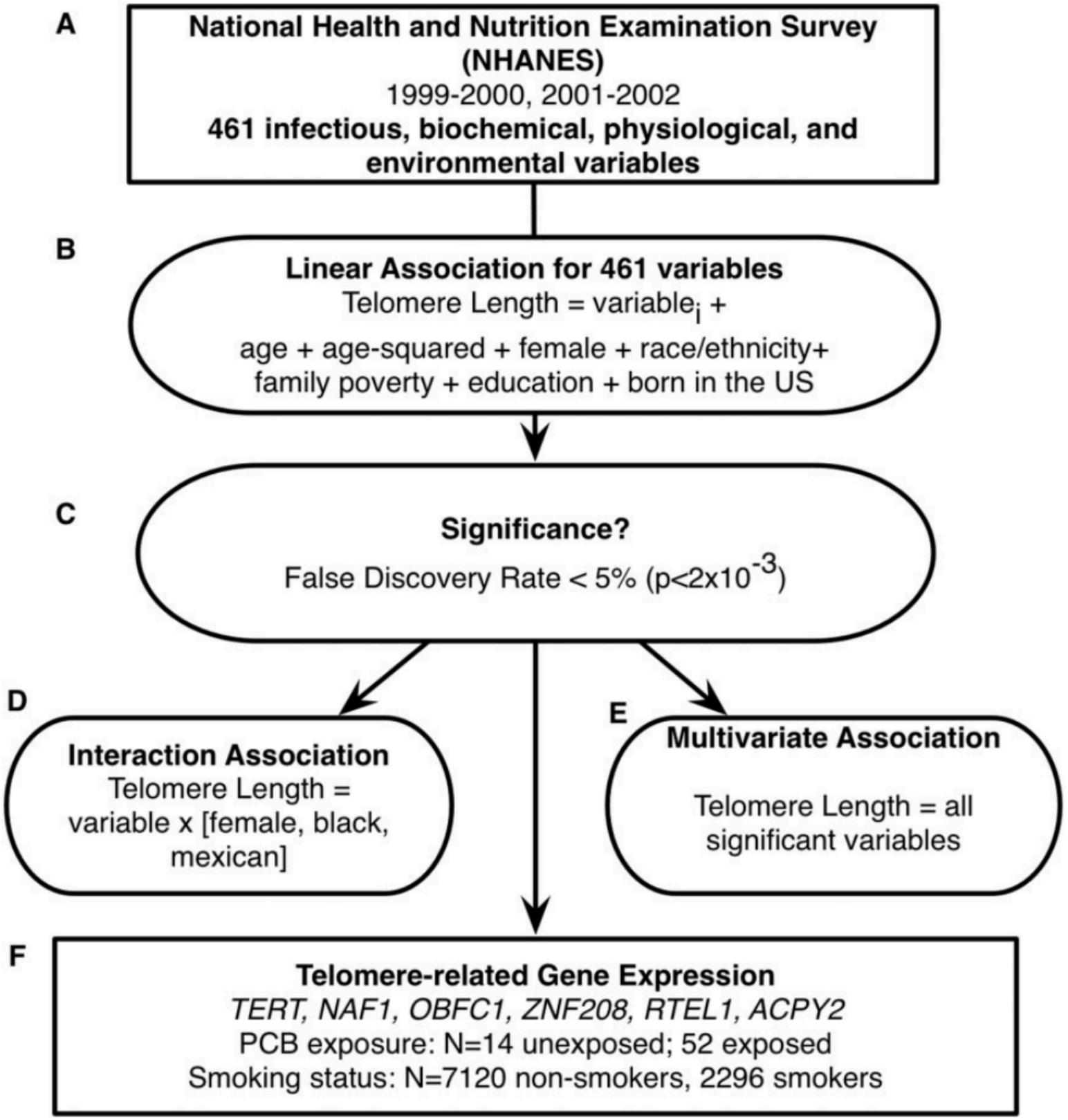
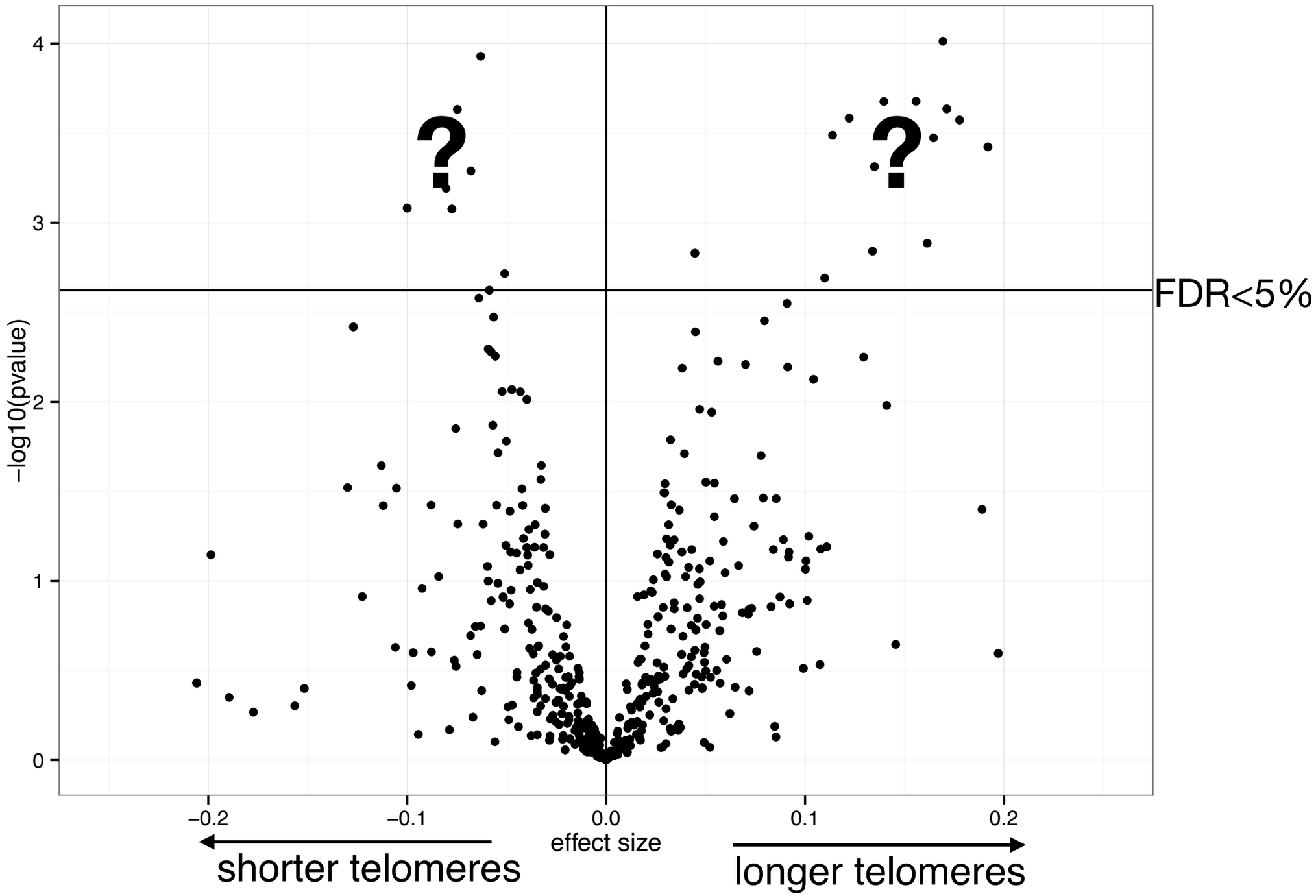


Figure 1. Method to search for physiological, environmental, and behaviour factors associated with mean telomere length (MTL).A) National Health and Nutrition Examination Survey of years 1999–2002. B) Scanning 461 variables for association iteratively in MTL. *i* denotes *i*th variable (out of 461) associated with MTL. C) Ascertaining statistical significance via false discovery rate. D) Interaction testing of FDR-significant variables with female sex, black, and Mexican race/ethnicity groups. E) Multivariate linear model predicting MTL as a function of FDR-significant variables. F) Estimating correlations between expression in genes that modulate telomere length in tissue samples exposed to PCB or smoking.

Associations in ***Telomere Length***:
Can you identify the associations in this graph?



median N=3000; N range: 300-7000

IJE, 2016

Recall:
'pseudo-code' for implementation of an *XWAS:*
how would you do it?

```
y = [blood pressure values for cohort]

association_list = empty_list()

for each x in list of exposures:
   association_test=f(x,y)
   append(association_list, association_test)

multiplicity_correct(association_list)

volcano_plot(y, x, association_list)
```

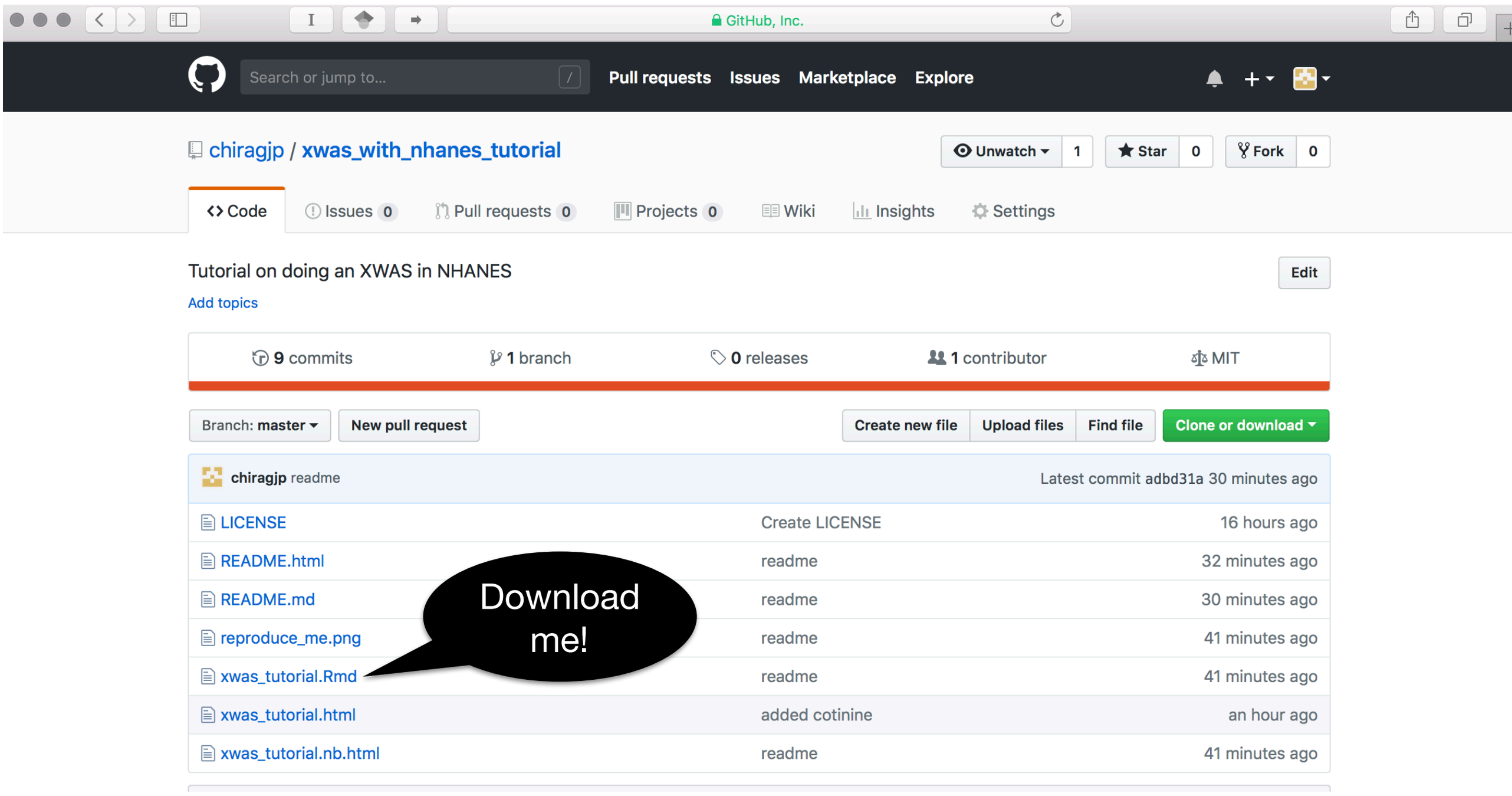
What is stored in *y*?
What is stored in *association_list*?
What can *f* be?
What does *append* do?
What is the reason for *multiplicity_correct?*

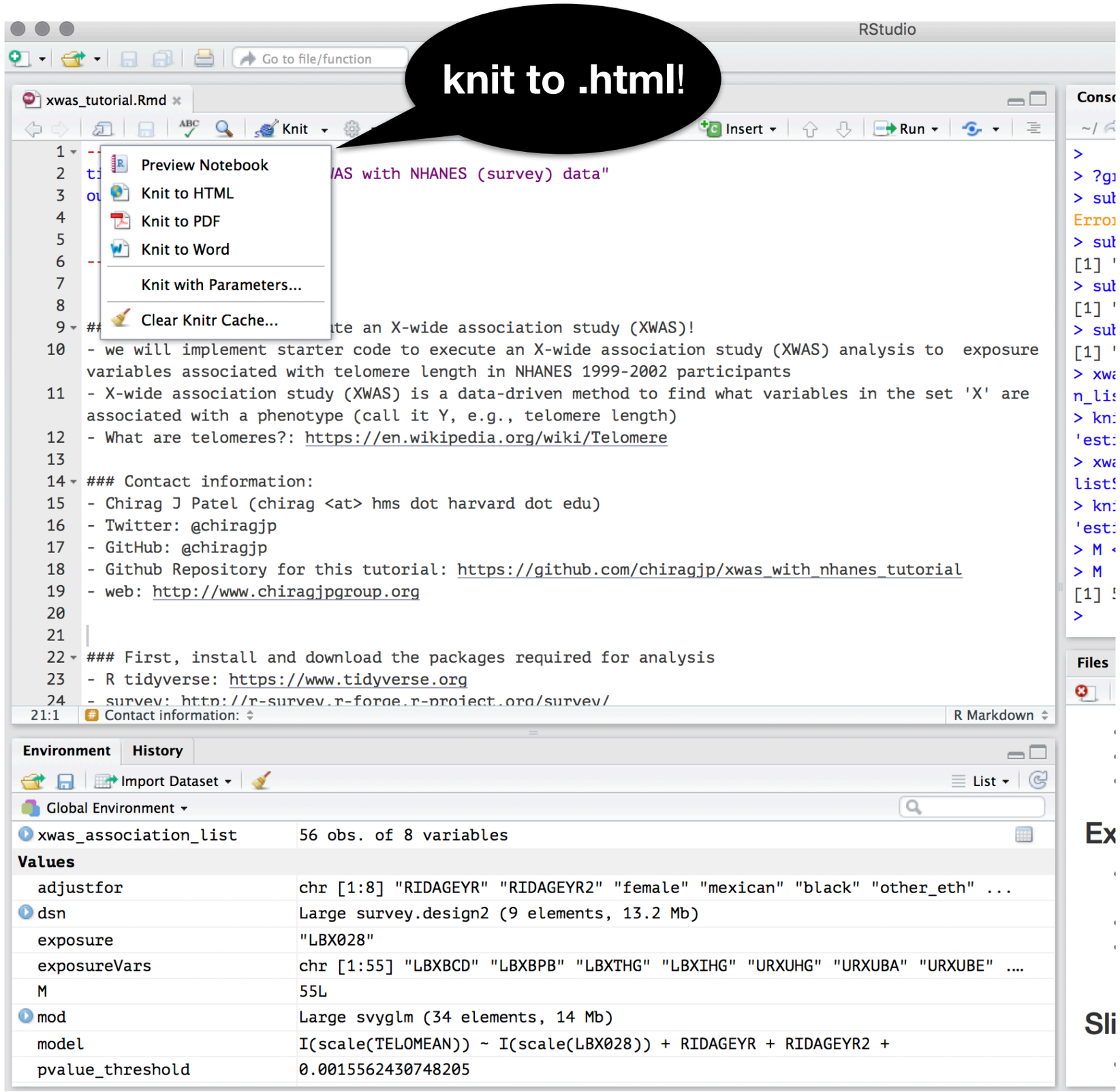
Manrai et al ARPH 2017

ok… lets get going coding!

http://bit.ly/xwas_with_nhanes

http://bit.ly/xwas_with_nhanes




RStudio
knit to .html!
Go to file/function
xwas_tutorial.Rmd
Knit
Insert
Run
Preview Notebook
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Clear Knitr Cache...
1 --
2 ti AS with NHANES (survey) data"
3 ou
4
5
6 --
7
8
9 ## te an X-wide association study (XWAS)!
10 - we will implement starter code to execute an X-wide association study (XWAS) analysis to exposure variables associated with telomere length in NHANES 1999-2002 participants
11 - X-wide association study (XWAS) is a data-driven method to find what variables in the set 'X' are associated with a phenotype (call it Y, e.g., telomere length)
12 - What are telomeres?: https://en.wikipedia.org/wiki/Telomere
13
14 ### Contact information:
15 - Chirag J Patel (chirag <at> hms dot harvard dot edu)
16 - Twitter: @chiragjp
17 - GitHub: @chiragjp
18 - Github Repository for this tutorial: https://github.com/chiragjp/xwas_with_nhanes_tutorial
19 - web: http://www.chiragjpgroup.org
20
21
22 ### First, install and download the packages required for analysis
23 - R tidyverse: https://www.tidyverse.org
24 - survey: http://r-survey.r-forge.r-project.org/survey/
21:1
Contact information:
R Markdown
Environment
History
Import Dataset
List
Global Environment
xwas_association_list
56 obs. of 8 variables
Values
adjustfor
chr [1:8] "RIDAGEYR" "RIDAGEYR2" "female" "mexican" "black" "other_eth" ...
dsn
Large survey.design2 (9 elements, 13.2 Mb)
exposure
"LBX028"
exposureVars
chr [1:55] "LBXBCD" "LBXBPB" "LBXTHG" "LBXIHG" "URXUHG" "URXUBA" "URXUBE" ....
M
55L
mod
Large svyglm (34 elements, 14 Mb)
model
I(scale(TELOMEAN)) ~ I(scale(LBX028)) + RIDAGEYR + RIDAGEYR2 +
pvalue_threshold
0.0015562430748205
Files

Questions

send a ‘pull’ request in GitHub so we can incorporate your answers!

- 0. What is your interpretation of the association sizes and pvalues for your XWAS?
- 1. How much to the coefficients change for the adjustment variables for each of the correlations?
- 2. Attempt to reproduce the findings in Patel et al., IJE 2016 (https://www.ncbi.nlm.nih.gov/pubmed/27059547) using more categories of exposure. How will you handle other ***X*** variables, such as self-reported variables?
- 3. When increasing the number of variables, how would the pvalue threshold change to accommodate more tests? How would the FDR change, if at all?
- 4. Execute the XWAS in another phenotype. What are the similarities and differences between your analysis in 1.
- 5. How much variance explained in telomeres do the top factors explain? Is this to be expected?
- 6. Implement the XWAS without using the `for` operator using the ***tidyverse*** suite of commands.

Acknowledgements

RagGroup
Nam Pho
Arjun Manrai
Jake Chung
Chirag Lakhani
Danielle Rasooly
Grace Mahoney

Harvard DBMI
Isaac Kohane

Stanford
John PA Ioannidis

NIEHS R00 ES023504
NIEHS R21 ES025052
NIAID R01 AI127250
NSF 1636870

DEPARTMENT OF
Biomedical Informatics

Chirag J Patel
chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org