

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

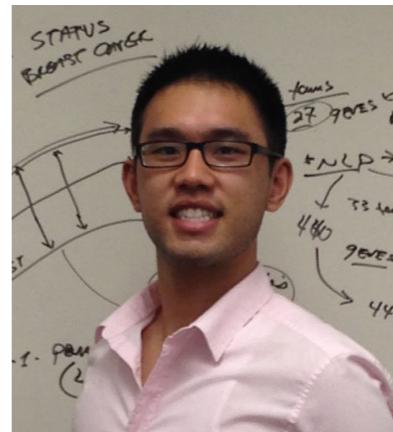
Please let us know if you are using the resources
(or provide feedback)!

Chirag



  @chiragjp

Nam



  @nampho2

Software prerequisites

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

http://bit.ly/xwas_with_nhances

The screenshot shows a GitHub repository page. At the top, the URL 'GitHub, Inc.' is visible in the address bar. The repository name 'chiragjp / xwas_with_nhances_tutorial' is displayed, along with 'Unwatch 1', 'Star 0', and 'Fork 0'. Below the repository name, there are tabs for 'Code', 'Issues 0', 'Pull requests 0', 'Projects 0', 'Wiki', 'Insights', and 'Settings'. The 'Code' tab is selected. A title 'Tutorial on doing an XWAS in NHANES' is present, with an 'Edit' button. There is also an 'Add topics' link. Below this, a summary bar shows '9 commits', '1 branch', '0 releases', '1 contributor', and 'MIT' license. A dropdown for 'Branch: master' and a 'New pull request' button are also shown. On the right, there is a 'Clone or download' button. The main area displays a list of files and their commit history. A large black speech bubble with white text 'Download me!' is overlaid on the left side of the file list.

File	Commit Message	Time
chiragjp readme		Latest commit abbd31a 30 minutes ago
LICENSE	Create LICENSE	16 hours ago
README.html	readme	32 minutes ago
README.md	readme	30 minutes ago
reproduce_me.png	readme	41 minutes ago
xwasTutorial.Rmd	readme	41 minutes ago
xwasTutorial.html	added cotinine	an hour ago
xwasTutorial.nb.html	readme	41 minutes ago

knit to .html!

```

1 -> Preview Notebook
2 ti IAS with NHANES (survey) data"
3 ou Knit to HTML
4 Knit to PDF
5 Knit to Word
6 Knit with Parameters...
7 Clear Knitr Cache...
8
9 ## Create 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\_nhances\_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:

```

Environment **History**

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] "LBXB028" "LBXB029" "LBXB030" "LBXB031" "LBXB032" "LBXB033" ...
M	55L
mod	Large svyglm (34 elements, 14 Mb)
model	I(scale(TELOMEAN)) ~ I(scale(LBX028)) + RIDAGEYR + RIDAGEYR2 +
pvalue_threshold	0.0015562430748205

Console **R Markdown**

```

>
> ?grep
> sub('\\(.+\\)', '', 'Uranium, urine (ng/mL)')
Error: '\\(' is an unrecognized escape in character string starting "'\\("
> sub('.+', '', 'Uranium, urine (ng/mL)')
[1] ""
> sub('\\(.+\\)', '', 'Uranium, urine (ng/mL)')
[1] "Uranium, urine "
> sub(' \\(.+\\)', '', 'Uranium, urine (ng/mL)')
[1] "Uranium, urine"
> xwas_association_list$var_desc <- sub(' \\(.+\\)', '', xwas_association_list$var_desc)
> knitr::kable(subset(xwas_association_list, fdr < 0.05)[, c('var_desc',
'estimate', 'p.value', 'fdr')])
```

Files **Plots** **Packages** **Help** **Viewer**

Execution

- R (version 3.3.3 or greater)
- Rstudio (version 1.01 or greater)
- Packages: survey, tidyverse, knitr, and broom

Slides

- Part 1, Data analytics to enable X-wide Association Studies (XWASs): https://github.com/chiragjp/xwas_with_nhances_tutorial/blob/master/xwas_tutorial.Rmd
- Part 2, Introduction to the hands-on tutorial to get you started to executing XWASs: https://github.com/chiragjp/xwas_with_nhances_tutorial/blob/master/part_2.ipynb

Real quick:
What is the *exposome*? What is the *phenome*?

exposome

internal

lead (serum)

nutrients (serum)

infection (urine)

metabolome

external

geography

air pollution

income

phenome

function

expression

telomere length

metabolome

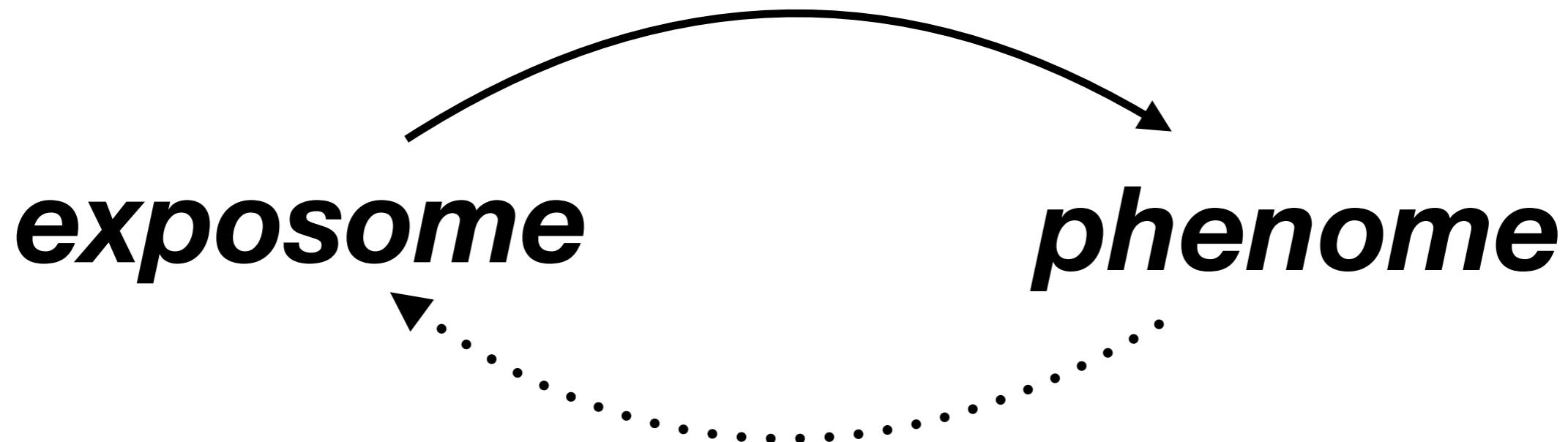
diseases

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 $\text{multiplicity_correct}$?

Gold standard for ***breadth*** of human exposure information: National Health and Nutrition Examination Survey¹



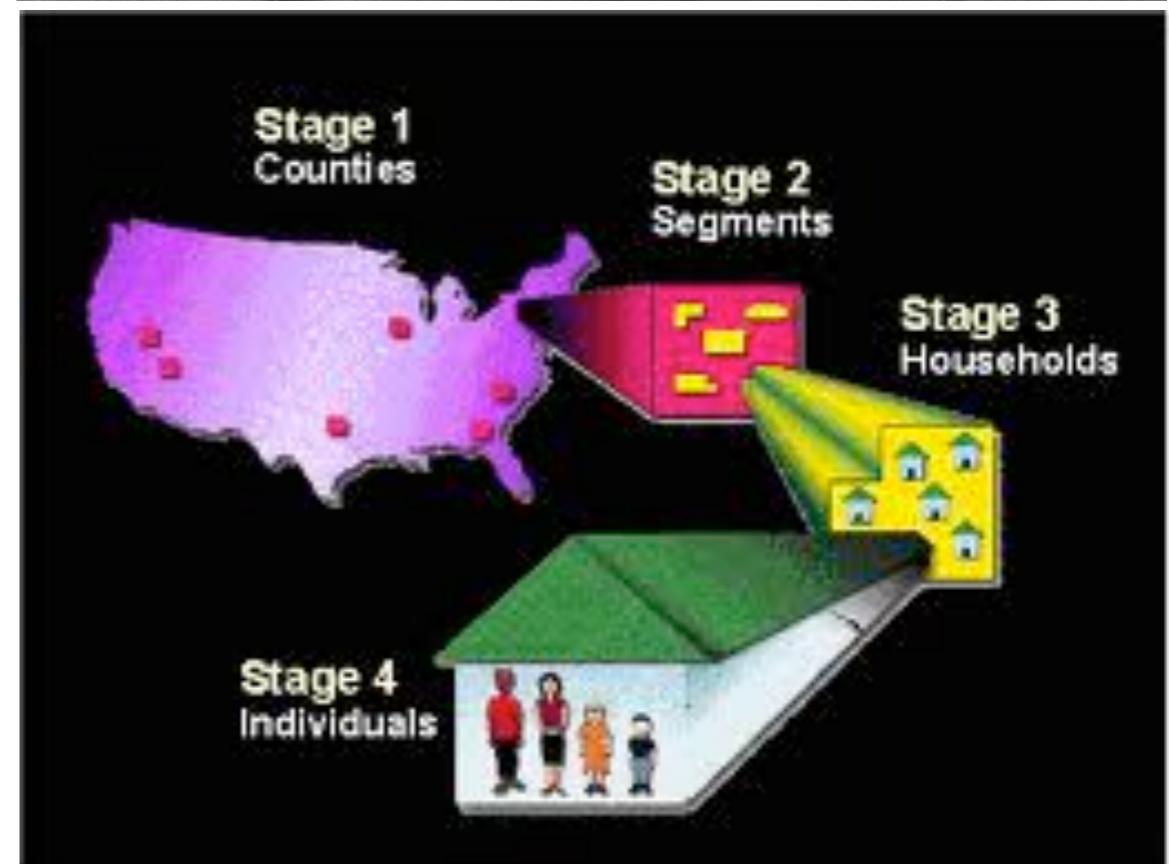
since the 1960s
now biannual: 1999 onwards
10,000 participants per survey



>250 exposures (serum + urine)
GWAS chip

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

Death index linkage (cause of death)

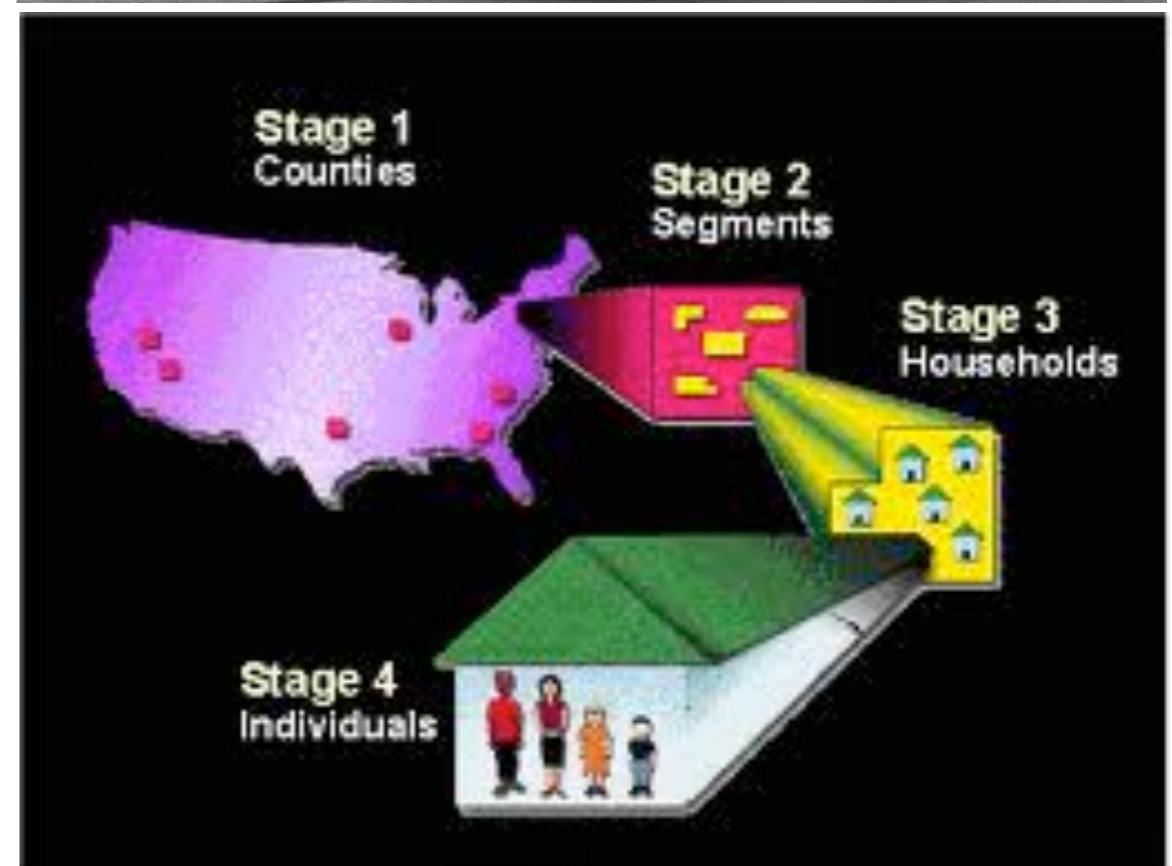


Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

Gold standard for *US health and phenotypic monitoring!*



- disease
 - heart disease
 - obesity
 - diabetes
- mortality

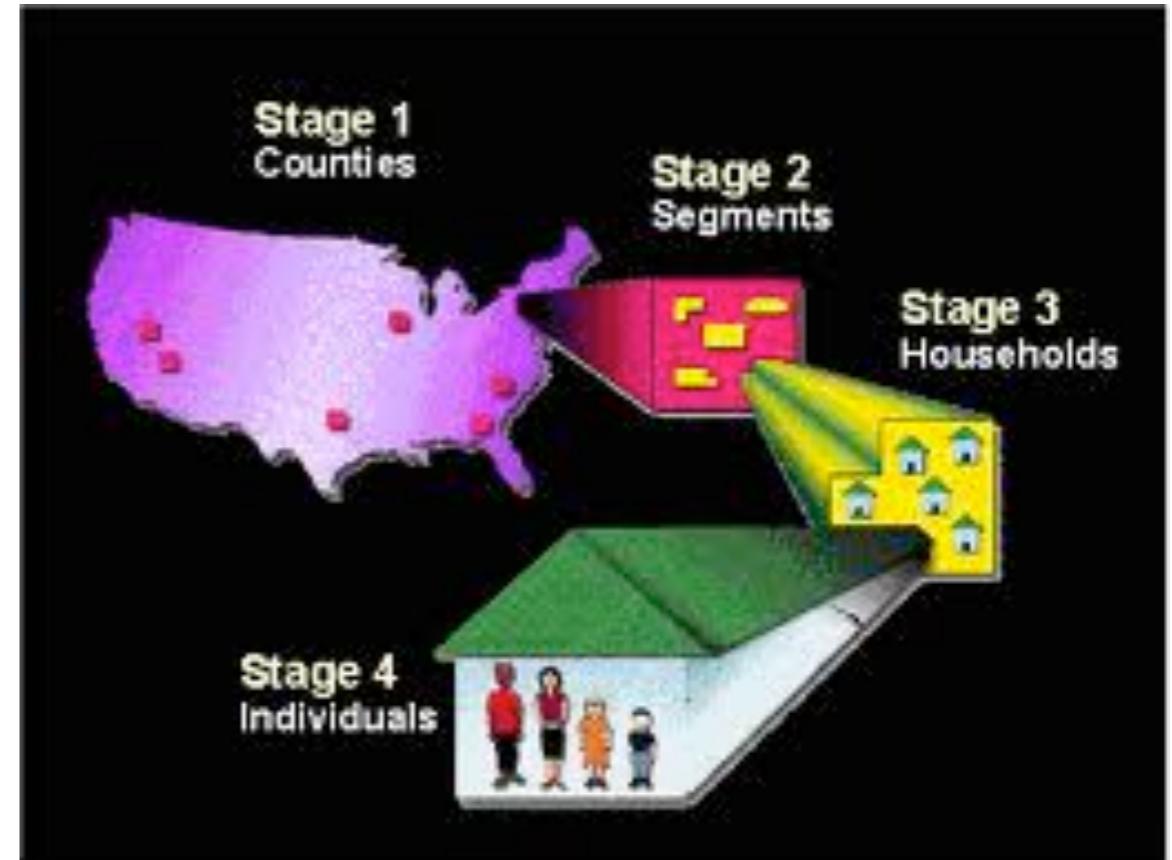


Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

Gold standard for *US health and phenotypic monitoring!*



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

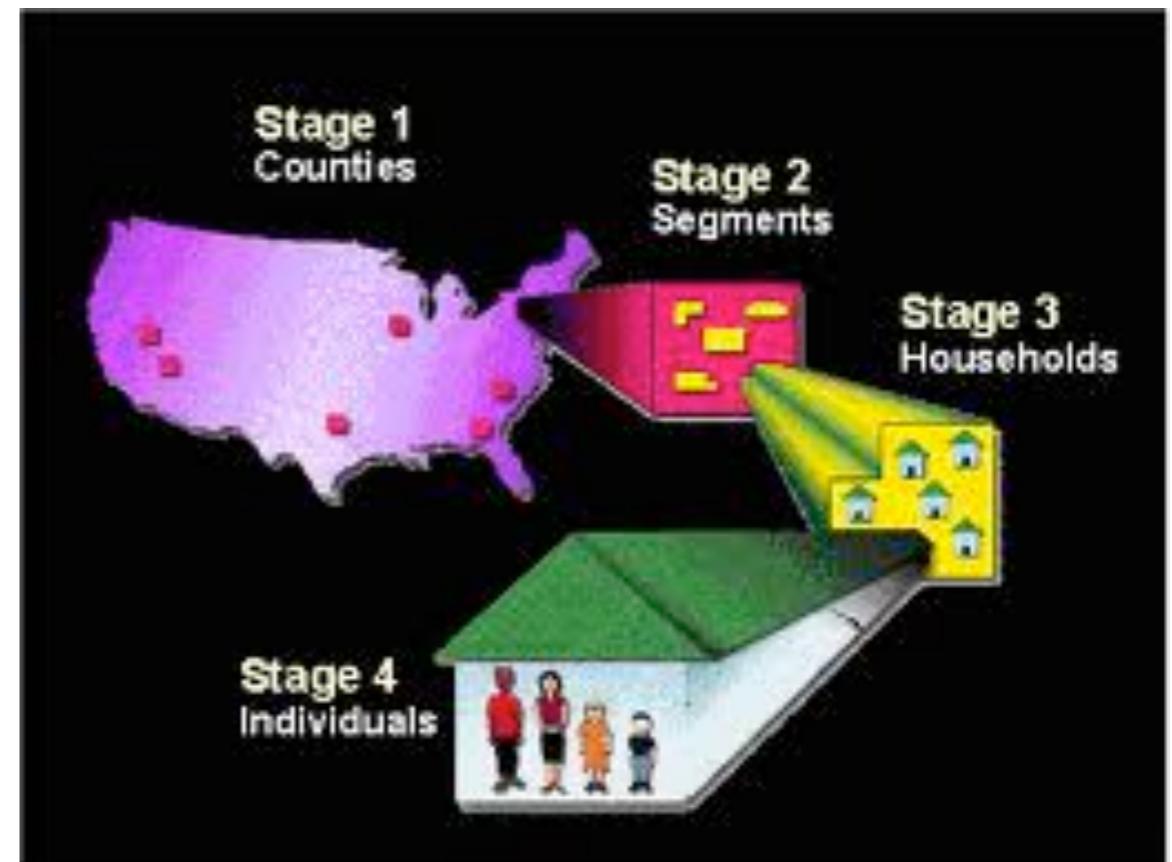


Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

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



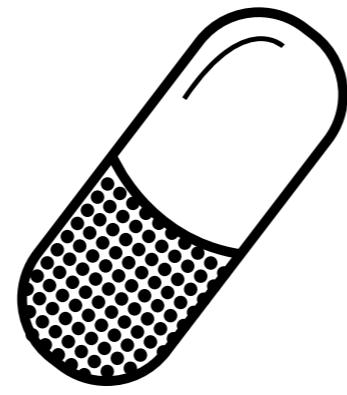
- behavior
 - diet
 - smoking
 - physical activity
- biomarkers of exposures



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

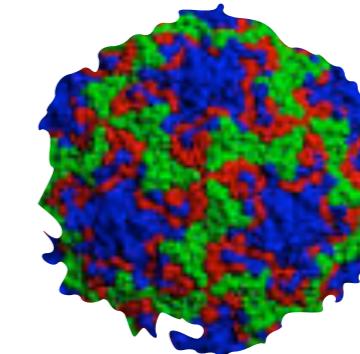


Nutrients and Vitamins
vitamin D, carotenes



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

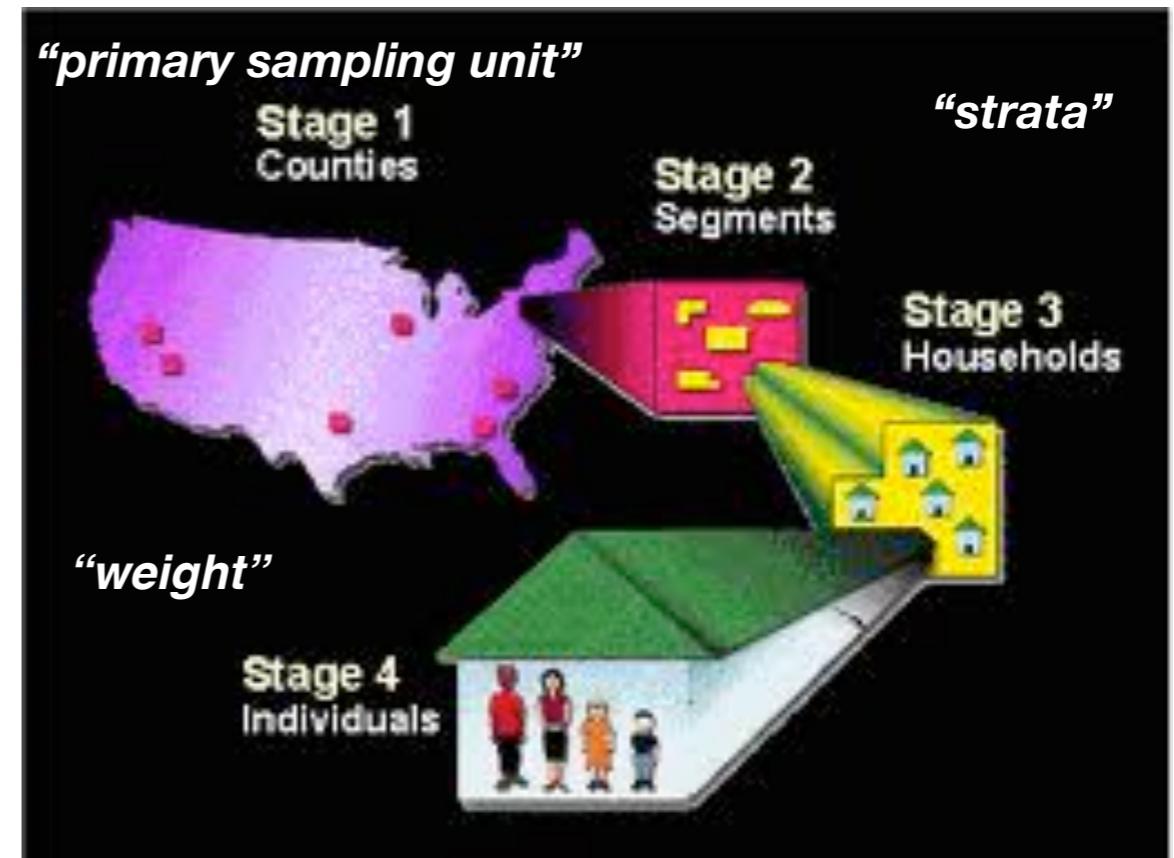
15 sampling units per year

- Counties
- Communities
- Households
- Individual

Over-“weights”:

- Seniors
- Children
- Ethnic minorities

“Stratified design”

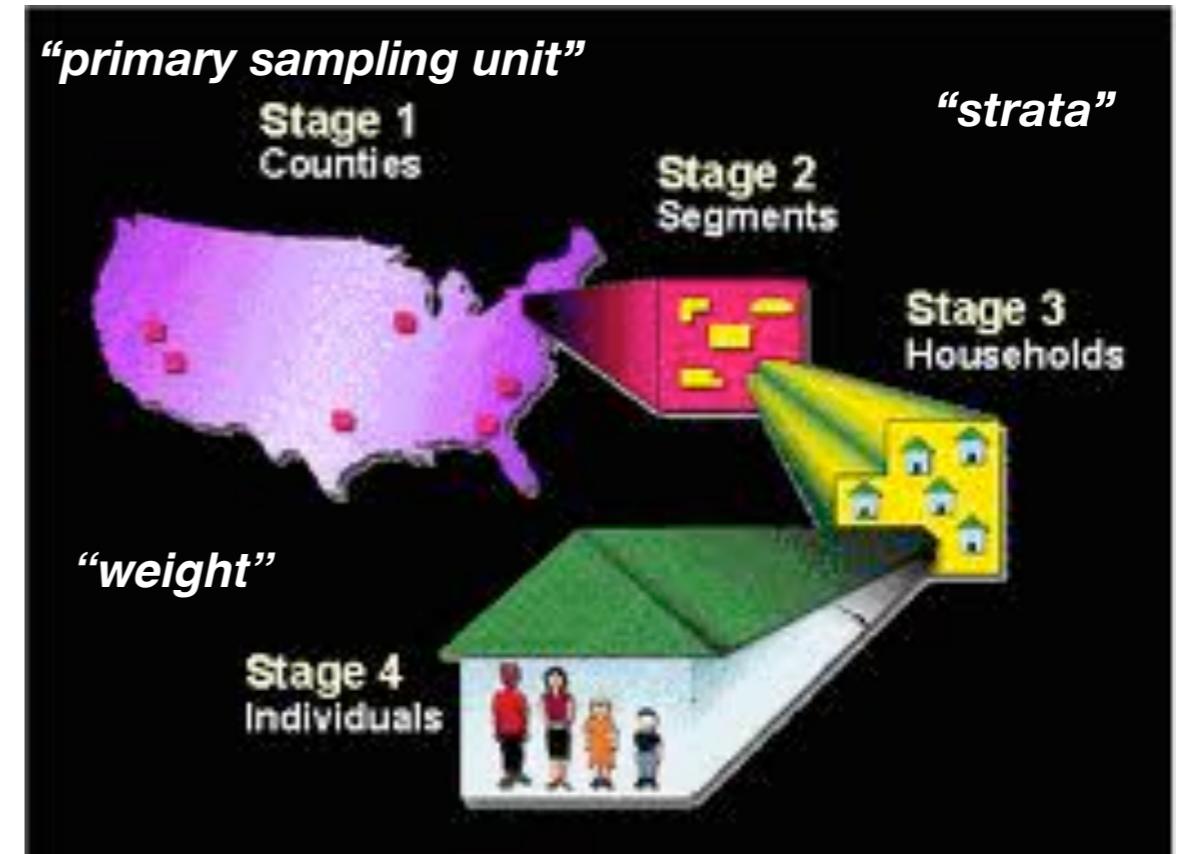




NHANES avoids sampling bias by executing a *stratified sampling design*

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?



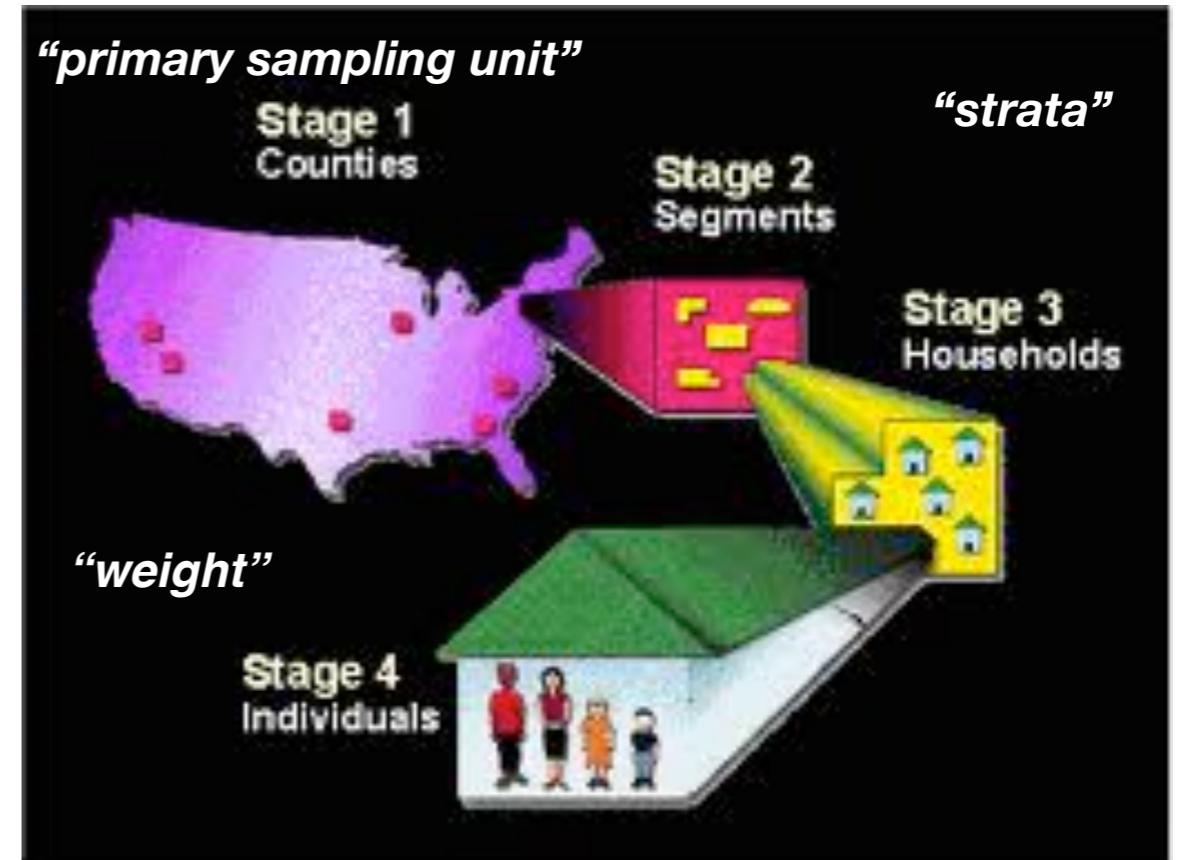
Recall: assumptions of regression and correlated data.



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



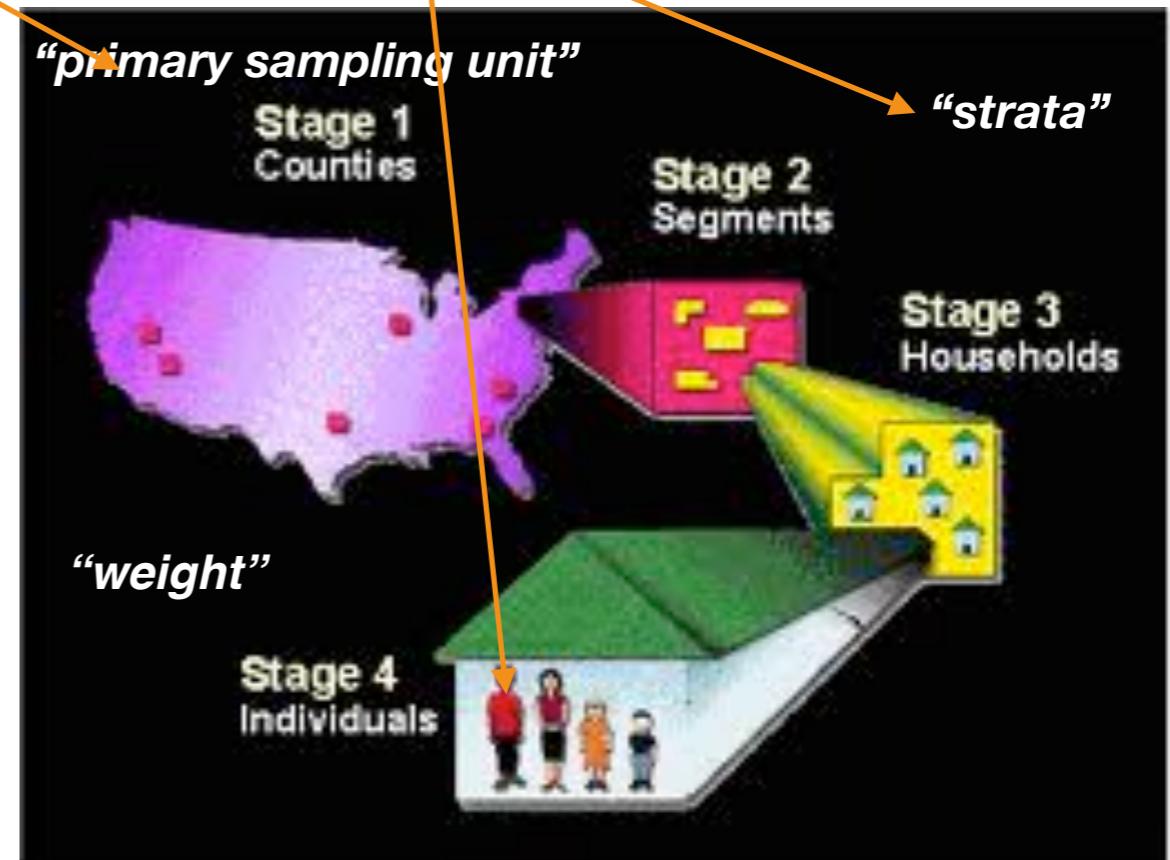
NHANES requires special analytic techniques that account for the stratified design: survey package in R



```
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.





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

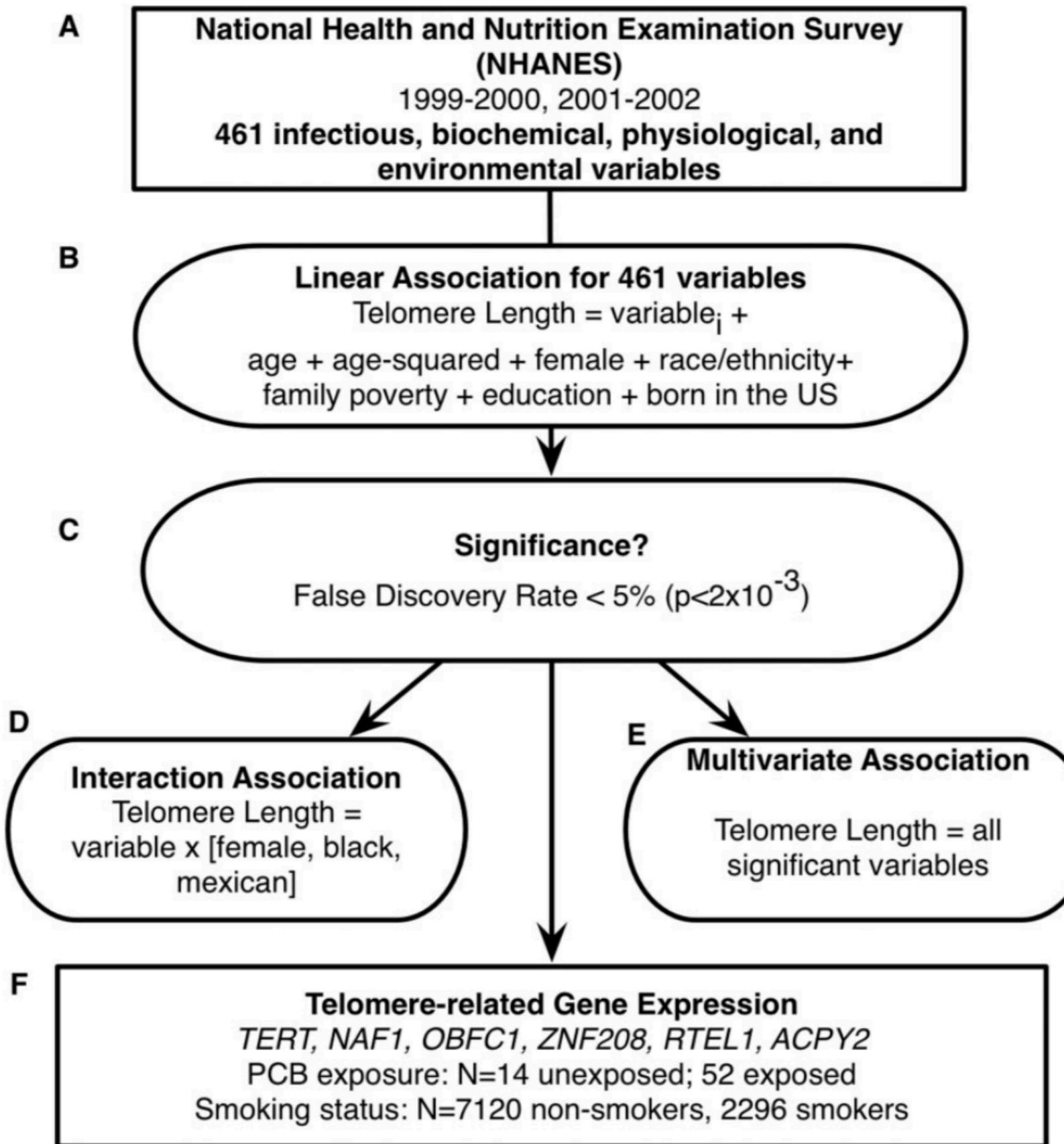
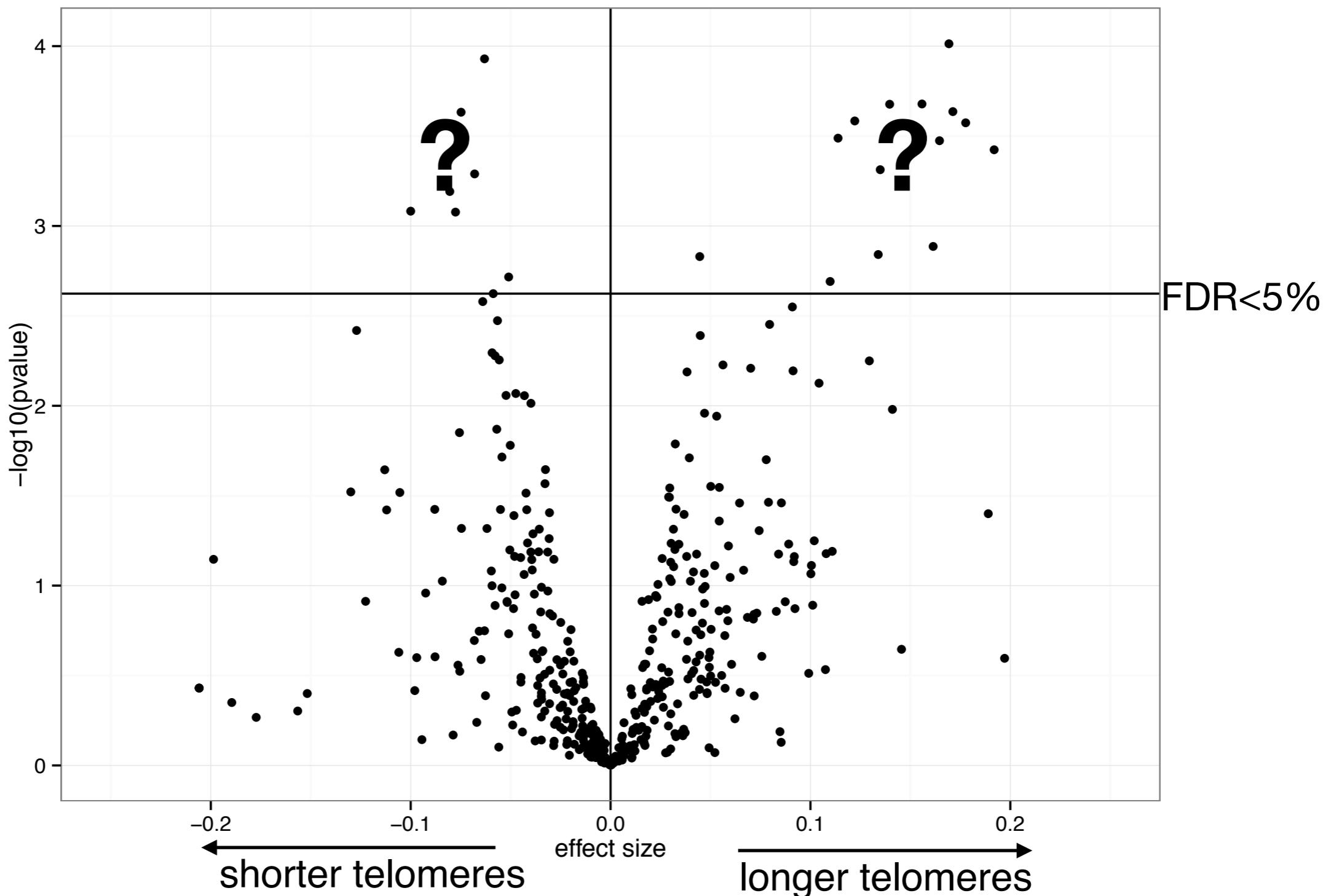


Figure 1. Method to search for physiological, environmental, and behaviour factors associated with mean telomere length (MTL). A) National Health and Nutrition Examination Survey of years 1999–2002. B) Scanning 461 variables for association iteratively in MTL. i denotes i th variable (out of 461) associated 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 $\text{multiplicity_correct}$?

ok... lets get going coding!

<http://bit.ly/xwas> with nhanes

http://bit.ly/xwas_with_nhances

The screenshot shows a GitHub repository page. At the top, the URL 'GitHub, Inc.' is visible in the address bar. The repository name 'chiragjp / xwas_with_nhances_tutorial' is displayed, along with 'Unwatch 1', 'Star 0', and 'Fork 0'. Below the repository name, there are tabs for 'Code', 'Issues 0', 'Pull requests 0', 'Projects 0', 'Wiki', 'Insights', and 'Settings'. The 'Code' tab is selected. A title 'Tutorial on doing an XWAS in NHANES' is present, with an 'Edit' button. There is also an 'Add topics' link. Below this, a summary bar shows '9 commits', '1 branch', '0 releases', '1 contributor', and 'MIT' license. A dropdown menu shows 'Branch: master' and a 'New pull request' button. To the right, there are buttons for 'Create new file', 'Upload files', 'Find file', and a green 'Clone or download' button. The main area lists files: 'chiragjp readme', 'LICENSE', 'README.html', 'README.md', 'reproduce_me.png', 'xwas Tutorial.Rmd', 'xwas Tutorial.html', and 'xwas Tutorial.nb.html'. A large black speech bubble with white text 'Download me!' is overlaid on the left side of the file list.

File	Commit Message	Time
chiragjp readme		Latest commit abbd31a 30 minutes ago
LICENSE	Create LICENSE	16 hours ago
README.html	readme	32 minutes ago
README.md	readme	30 minutes ago
reproduce_me.png	readme	41 minutes ago
xwas Tutorial.Rmd	readme	41 minutes ago
xwas Tutorial.html	added cotinine	an hour ago
xwas Tutorial.nb.html	readme	41 minutes ago

knit to .html!

```

1 -> tis "X-wide association study (XWAS)!" 
2 -> ouas "XWAS with NHANES (survey) data"
3 -> Knit to HTML
4 -> Knit to PDF
5 -> Knit to Word
6 -> Knit with Parameters...
7 -> Clear Knitr Cache...
8 -> # Create an X-wide association study (XWAS)!
9 -> # we will implement starter code to execute an X-wide association study (XWAS) analysis to exposure
10 -> 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
12 -> associated with a phenotype (call it Y, e.g., telomere length)
13 -> - What are telomeres?: https://en.wikipedia.org/wiki/Telomere
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\_nhances\_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:

```

Environment **History**

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] "LBXB028" "LBXB029" "LBXB030" "LBXB031" "LBXB032" "LBXB033" ...
M	55L
mod	Large svyglm (34 elements, 14 Mb)
model	I(scale(TELOMEAN)) ~ I(scale(LBX028)) + RIDAGEYR + RIDAGEYR2 +
pvalue_threshold	0.0015562430748205

Console **R Markdown**

```

>
> ?grep
> sub('\\(.+\\)', '', 'Uranium, urine (ng/mL)')
Error: '\\(' is an unrecognized escape in character string starting "'\\("
> sub('.+', '', 'Uranium, urine (ng/mL)')
[1] ""
> sub('\\(.+\\)', '', 'Uranium, urine (ng/mL)')
[1] "Uranium, urine "
> sub(' \\(.+\\)', '', 'Uranium, urine (ng/mL)')
[1] "Uranium, urine"
> xwas_association_list$var_desc <- sub(' \\(.+\\)', '', xwas_association_list$var_desc)
> knitr::kable(subset(xwas_association_list, fdr < 0.05)[, c('var_desc',
  'estimate', 'p.value', 'fdr')])
```

Files **Plots** **Packages** **Help** **Viewer** **Publish**

- R (version 3.3.3 or greater)
- Rstudio (version 1.01 or greater)
- Packages: survey, tidyverse, knitr, and broom

Execution

- download 'xwas_tutorial.Rmd' in RStudio from here: https://github.com/chiragjp/xwas_with_nhances_tutorial
- knit the .Rmd in RStudio
- or download a pre-knitted file: https://github.com/chiragjp/xwas_with_nhances_tutorial/blob/master/xwas_tutorial.html

Slides

- Part 1, Data analytics to enable X-wide Association Studies (XWASs): https://github.com/chiragjp/xwas_with_nhances_tutorial/blob/master/part1.Rmd
- Part 2, Introduction to the hands-on tutorial to get you started to executing XWASs: https://github.com/chiragjp/xwas_with_nhances_tutorial/blob/master/part2.Rmd