# Leveraging Longitudinal and Biological Information in Genetic Epidemiology Research

Peng Wei, Ph.D.
Department of Biostatistics &
Human Genetics Center

University of Texas School of Public Health
October 26th, 2015

# Introduction

- Many prospective cohort studies and electronic health record (EHR)-based cohorts have collected longitudinal exposure and phenotype information
  - **ARIC** study: n = ~11,000 EA and ~4,000 AA
  - Genetic Epidemiology Research on Adult Health and Aging (**GERA**) Cohort: n = 100,000
  - **UK Biobank**: n = 500,000
  - **Precision Medicine Initiative** national research cohort (planned): n > 1,000,000
- However, most genetic epidemiology research has thus far largely ignored the rich longitudinal information and simply used the baseline measurements to identify genotype-phenotype associations
- New and powerful statistical methods for analyzing longitudinal data in genetic epidemiology research are needed

# Part I:
# Gene x Longitudinal Exposure Interaction

RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

# Functional Logistic Regression Approach to Detecting Gene by Longitudinal Environmental Exposure Interaction in a Case-Control Study

Peng Wei,[1]* Hongwei Tang,[2] and Donghui Li[2]

[1] *Division of Biostatistics and Human Genetics Center, The University of Texas School of Public Health, Houston, Texas, United States of America;*
[2] *Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Genet Epidemiol 38:638–651, 2014.

# A motivating example: pancreatic cancer as a complex disease

- The 4th leading cause of cancer-related deaths in both men and women with a 5-year survival rate of 6%

## Genetic risk factors

- ABO blood group (PanScan I GWAS, 2009)
- NR5A2, CLPTM1L-TERT, HNF1A (PanScan II, 2010)
- LINC-PINT, BCAR1, PDX1, ZNRF3 (PanScan III, 2014)
- Biological pathways (Wei et al, PLoS ONE 2012):
  - Maturity onset diabetes of the young (MODY)
  - G protein-coupled receptor signaling pathway

## Epi risk factors

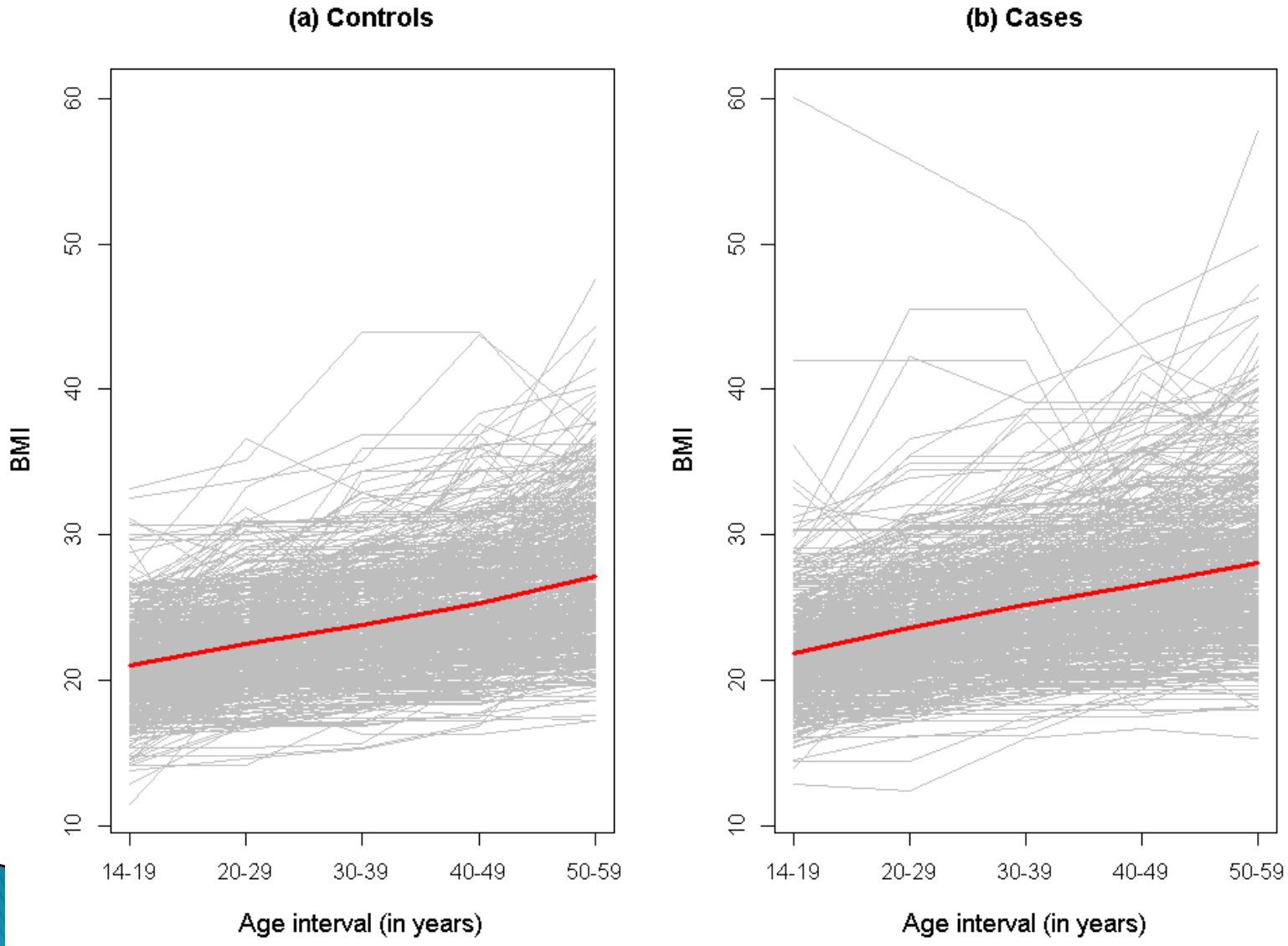
- Cigarette smoking
- Family history
- Long-term type 2 diabetes
- Heavy alcohol consumption
- Obesity (weight loss/cachexia may be early symptom-reverse causality)

**Gene-Environment Interaction (GxE)**

# Gene x BMI interaction as an example of GxE

- "E" here: body mass index (BMI) defined as body weight (kg)/body height $(m)^2$
- Most existing GxE methods consider E as static and measured at some arbitrary time point
- BMI changes across one's lifespan
- Which time window's BMI should be used?
  - BMI at diagnosis – bad idea (reverse-causality)
  - Usual adulthood BMI – maybe, but adulthood means 30-40 years
  - Lifespan-BMI profile – sounds reasonable, but how?
- Like BMI, many environmental exposures/host factors change over time, e.g., diet, air pollution, and cholesterol level
- We need new statistical methods and analysis strategies to deal with longitudinal environmental factors in GxE analysis

# A motivating example:
# UT MD Anderson Case Control Study of Pancreatic Cancer




Li *et al*, JAMA 2009

# Standard GxE model for a binary trait

$$\text{logit}[\Pr(D=1|X,G,E)] = \log \frac{\Pr(D=1|X,G,E)}{1-\Pr(D=1|X,G,E)}$$

$$= \alpha_0 + \alpha X + \beta_G G + \beta_E E + \beta_{GE} G \times E$$

Covariates

SNP

Environmental factor

Test $H_0$: $\beta_{GE}=0$

# Gene x Longitudinal BMI Interaction –Functional Logistic Regression (FLR)

- First, we use FLR to model the **main effect** of the longitudinal BMI
- We assume that the disease status is associated with the longitudinal BMI via the following functional logistic regression model,

Time-varying coefficient

$$\text{logit}[\Pr(D_i=1)]=\alpha_0+\int_{\mathcal{T}}\beta(t)E_i(t)dt\,,$$

where **the longitudinal BMI follows a stochastic process** $\{E(t), t\in\mathcal{T}\}$, which has mean function $\mu(t)$, and covariance function $R(s,t)=\text{cov}\{E(s),E(t)\}$.

- Interpretation: a constant unit increase in BMI from time $t_1$ to $t_2$ will increase the log odds of disease by $\int_{t_1}^{t_2}\beta(t)\,dt$.

# Gene x Longitudinal BMI Interaction– Functional Principal Component Analysis (FPCA) of longitudinal exposure

- We have eigen-decomposition $R(s,t) = \sum_{k=1}^{\infty} \lambda_k \rho_k(s)\rho_k(t)$, where $\rho_k$ and $\lambda_k$ are eigenfunctions and eigenvalues ordered by size $\lambda_1 \geq \lambda_2 \geq \cdots$
- By the *Karhunen-Loève* theorem, $E_i(t)$ can be represented by $\boldsymbol{E_i(t) = \sum_{k=1}^{\infty} FPC_{ik}\rho_k(t)}$, where $FPC_{ik} = \int E_i(t)\rho_k(t)dt$ is the $k^{th}$ functional principal component (FPC) score for the $i^{th}$ subject
- E($FPC_k$)=0, Var($FPC_k$) = $\lambda_k$, cov($FPC_k$, $FPC_j$) = 0 if $k \neq j$
- Thus **each $\boldsymbol{E_i(t)}$, i.e., BMI profile, is represented as a sum of orthogonal curves with uncorrelated random coefficients $\boldsymbol{FPC_{ik}}$**, which measures the "similarity" between $E_i(t)$ and the *k*th eigenfunction ("pattern") $\rho_k(t)$
- Time-varying coefficient $\boldsymbol{\beta(t) = \sum_{k=1}^{\infty} \beta_k \rho_k(t)}$.

# Gene x Longitudinal BMI Interaction – FLR approach

- From the orthonormality of the eigenfunctions $\rho_k$, it follows that $\int \beta(t)\, E_i(t) dt = \sum_{k=1}^{\infty} \beta_k\, FPC_{ik}$.
- We propose a truncated version of the FLR regression,

  $\text{logit}[\Pr(D_i = 1)] = \alpha_0 + \sum_{k=1}^{K} \beta_k\, FPC_{ik}$,

  i.e., **the outcome $D$ is dependent only on the leading $K$ FPCs**
- The eigenfunctions, FPC scores can be estimated from the sparsely observed longitudinal BMI data, e.g., via the PACE method (Yao et al, JASA 2005)

# FPCA by PACE

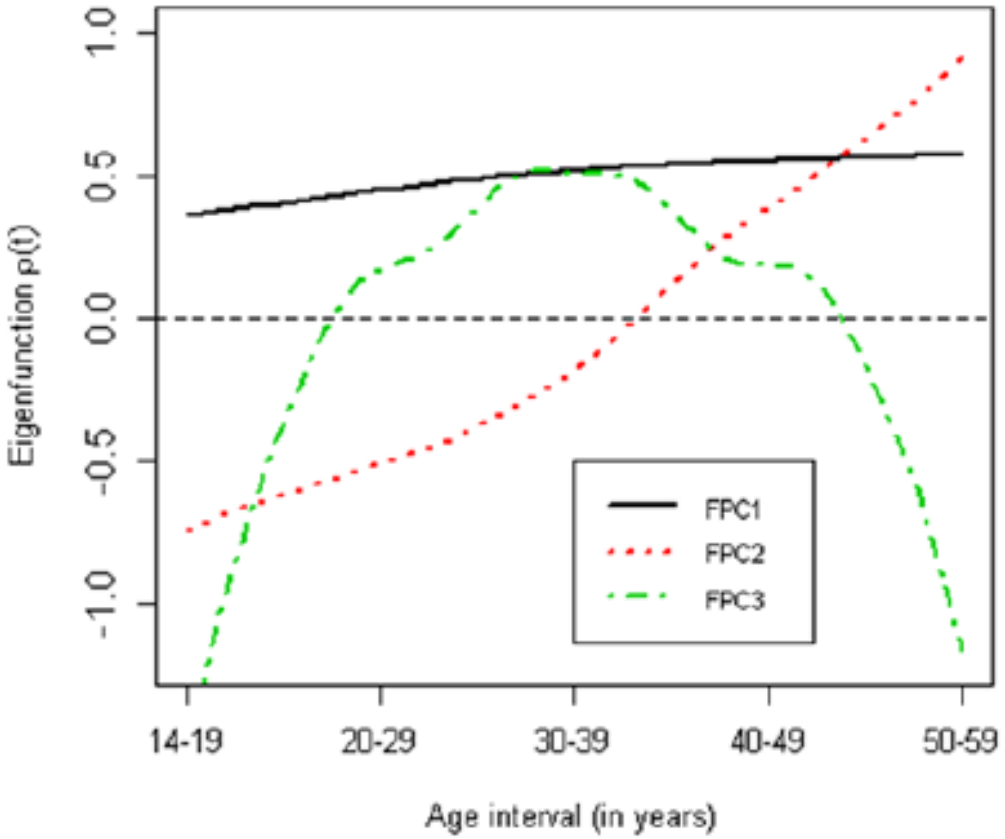
- PACE takes into account the measurement error in the observed longitudinal exposure trajectory (recalled lifetime body weights, thus, BMIs, in the example here)
- $BMI_{ij} = E_i(t_{ij}) + \varepsilon_{ij}$, where $\varepsilon_{ij}$ are zero-mean errors with variance $\sigma_\varepsilon^2$, and are independent of $E_i(.)$ – the classical measurement error model (Carroll et al, 2006)
- So the observed BMI is modeled as:

$$BMI_{ij} = \mu(t_{ij}) + \sum_{k=1}^{K} FPC_{ik}\, \rho_k(t_{ij}) + \varepsilon_{ij}$$

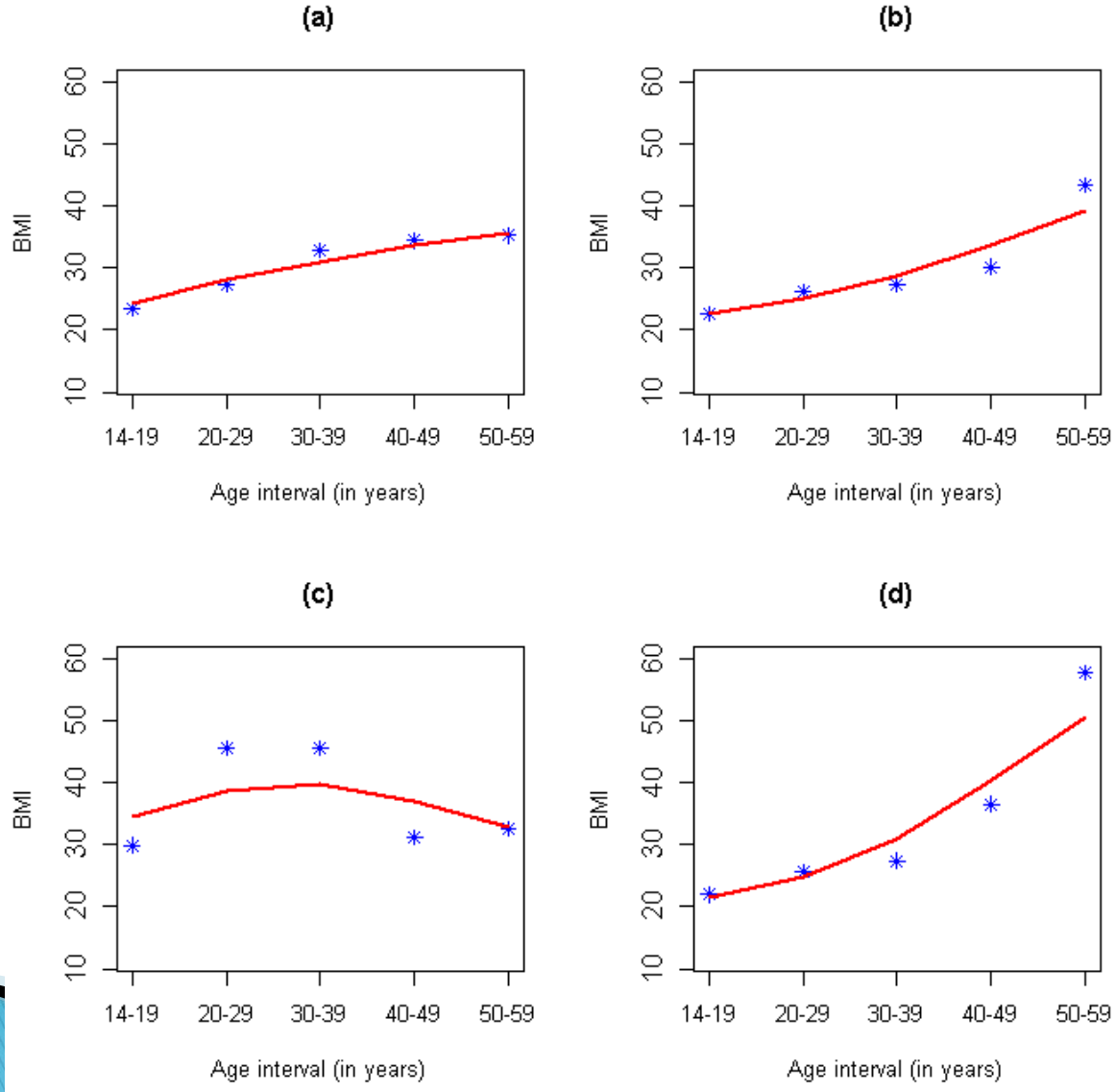
Estimated by PACE

Estimated by kernel smoothing

Eigen-decomposition of kernel smoothed estimated covariance function $\hat{R}(s,t)$


Eigenfunction ρ(t)
1.0
0.5
0.0
-0.5
-1.0
14-19
20-29
30-39
40-49
50-59
Age interval (in years)
FPC1
FPC2
FPC3

# 4 individuals: observed vs FPCA-fitted BMI profiles



# Gene x Longitudinal BMI Interaction – Functional Logistic Regression (FLR)

- By resorting to the FLR, we re-formulate

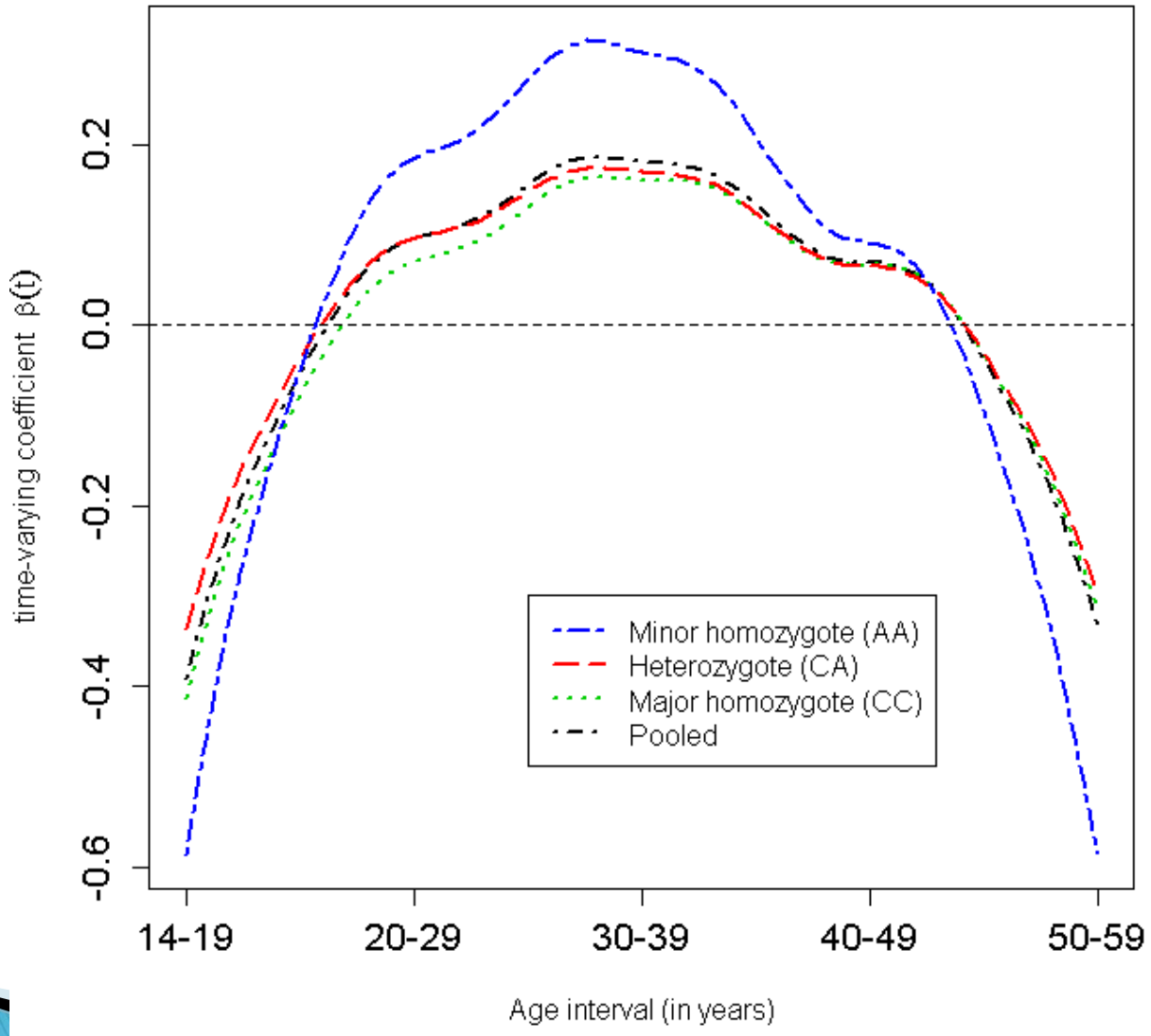
$$\text{logit}[\Pr(D_i=1)]=\alpha_0+\int \beta(t)\,E_i(t)dt$$

as: $\text{logit}[\Pr(D_i=1)]=\alpha_0+\sum_{k=1}^{K}\beta_k\,FPC_{ik}$,
where $FPC_{ik}=\int E_i(t)\rho_k(t)dt$ (by PACE)

- Then the estimated time-varying coefficient of BMI is $\widehat{\beta(t)}=\sum_{k=1}^{K}\widehat{\beta_k}\,\widehat{\rho_k(t)}$

| | Estimates | SE * | 95% CI* | SE** | 95% CI** | P-value* |
|---|---|---|---|---|---|---|
| $\beta_1$ | .06 | .01 | **(.04, .08)** | .009 | **(.04, .08)** | **1.39e-09** |
| $\beta_2$ | -.04 | .03 | (-.10, .02) | .03 | (-.10, .02) | .15 |
| $\beta_3$ | .28 | .14 | (.003, .56) | .15 | (-.03, .58) | **.048** |

* Model-based; **Nonparametric bootstrap


time-varying coefficient β(t)
0.2
0.0
-0.2
-0.4
-0.6
14-19
20-29
30-39
40-49
50-59
Age interval (in years)
Minor homozygote (AA)
Heterozygote (CA)
Major homozygote (CC)
Pooled

# Gene x Longitudinal BMI Interaction – FLR approach

- For gene x longitudinal BMI interaction, the proposed FLR is:

Indicator variable for SNP

$$\text{logit}[\Pr(D_i=1)]=\alpha_0+\beta_G G_i+\sum_{k=1}^{K}\beta_k FPC_{ik}+\sum_{k=1}^{K}\beta_{Gk}\, G_i * FPC_{ik}$$

Main effects for longitudinal BMI

Interaction effects

- We test $H_0: \beta_{G1}=\beta_{G2}=\cdots=\beta_{GK}=0$
- Likelihood ratio test (LRT) or score test: $\chi^2(K)$
- A powerful alternative: sum of squared score test (SSU)

# Simulation: Type I error ($\alpha = 0.05$)

| Meas. error in FPCA simu model | BMI (14–19 yrs.) | BMI (20–29 yrs.) | BMI (30–39 yrs.) | BMI (40–49 yrs.) | BMI (50–59 yrs.) | BMI MinP (unadjusted) | BMI MinP (Bonf.) | BMI MinP (Para. bootstrap) | BMI (Score) | FLR (Score) | FLR (SSU) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| w/o error | .051 | .049 | .048 | .047 | .054 | **.103** | .020 | .025 | .049 | .052 | .048 |
| with error ($\sigma$) | .047 | .049 | .047 | .047 | .049 | **.121** | .026 | .031 | .044 | .050 | .050 |
| with error ($2\sigma$) | .041 | .047 | .046 | .049 | .047 | **.140** | .031 | .038 | .043 | .051 | .044 |

*20,000 replications: 1,000 cases and 1000 controls in each replication

# Simulation: Power ($\alpha = 0.05$)

| Meas. error in FPCA simu model | BMI (14–19 yrs.) | BMI (20–29 yrs.) | BMI (30–39 yrs.) | BMI (40–49 yrs.) | BMI (50–59 yrs.) | BMI MinP (unadjusted) | BMI MinP (Bonf.) | BMI MinP (Para. bootstrap) | BMI (Score) | FLR (Score) | FLR (SSU) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| w/o error | .768 | .842 | .837 | .762 | .579 | .873* | .740 | .754 | .794 | .708 | **.811** |
| with error ($\sigma$) | .727 | .815 | .816 | .734 | .570 | .893* | .724 | .736 | .593 | .690 | **.810** |
| with error ($2\sigma$) | .633 | .739 | .754 | .658 | .509 | .894* | .711 | .719 | .564 | .655 | **.792** |

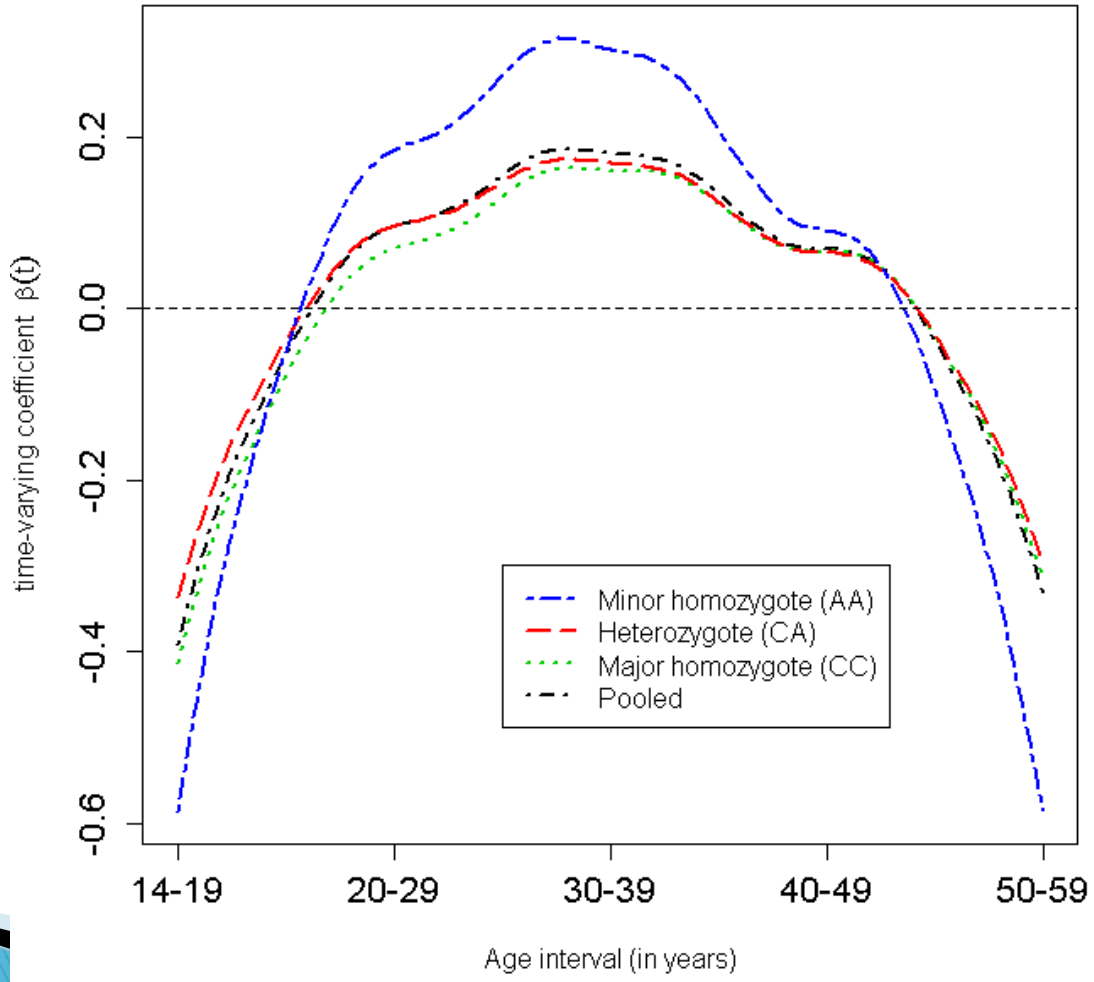
* Type I error cannot be controlled

# Gene x Longitudinal BMI Interaction – FLR

- We applied the proposed FLR method to the MD Anderson case-control study individuals with genotypes in candidate genes (553 cases and 580 controls)

| SNP | Gene | BMI (14-19) | BMI (20-29) | BMI (30-39) | BMI (40-49) | BMI (50-59) | BMI (Bonf-erroni) | BMI (MinP; bootstrap) | BMI (Score) | FLR (Score) | FLR (SSU) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs505922 | ABO | .63 | .28 | .87 | .82 | .67 | 1.0 | .83 | .42 | .61 | .77 |
| rs1558902 | FTO | .11 | .14 | .72 | .33 | .99 | .55 | .46 | .06 | .39 | .36 |
| **rs8050136** | **FTO** | **.04** | **.02** | **.04** | **.03** | .12 | .12 | .14 | .17 | .10 | **.02** |
| rs12029406 | NR5A2 | .69 | .98 | .15 | .09 | .19 | .44 | .38 | .23 | .39 | .18 |
| rs3790844 | NR5A2 | .62 | .73 | .35 | .09 | .11 | .43 | .34 | .63 | .43 | .23 |
| rs3790843 | NR5A2 | .58 | .96 | .22 | .13 | .21 | .65 | .54 | .70 | .65 | .32 |
| rs401681 | CLPTM1L-TERT | .42 | .99 | .94 | .95 | .98 | 1.0 | .94 | .88 | .63 | 1.0 |

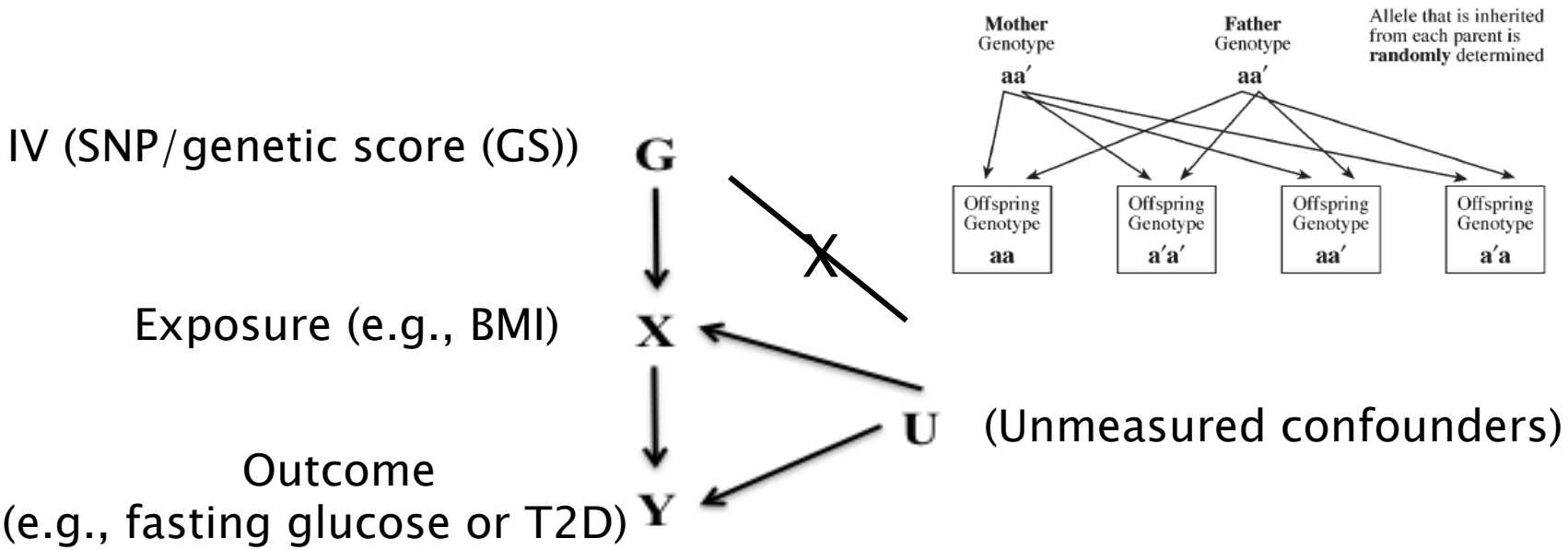
# Genotype-specific time-varying interaction coefficient for the longitudinal BMI: rs8050136 in FTO



# Part II:
# Mendelian randomization analysis with a time-varying exposure

# Mendelian randomization (MR) analysis

- MR is a type of instrumental variable (IV) analysis with SNPs as the IVs

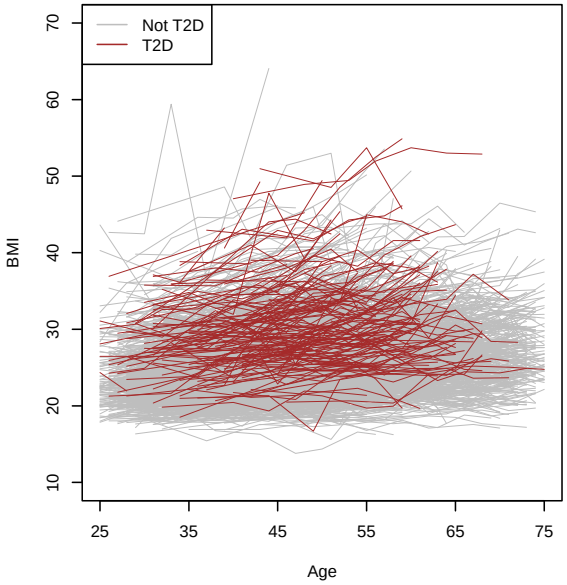


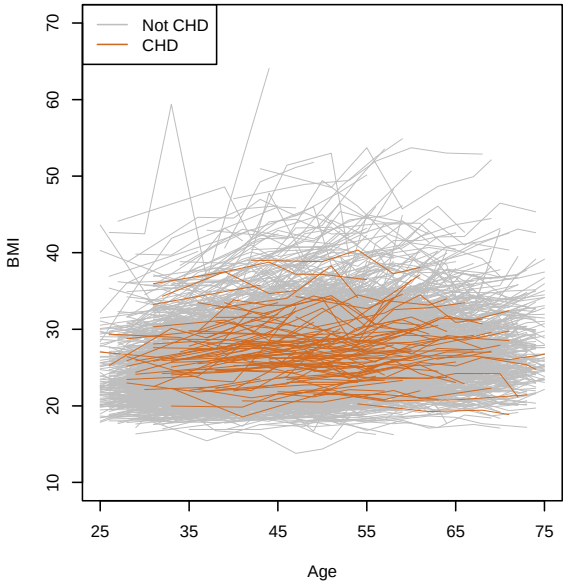

- Two-stage least squares (2SLS) for continuous Y:
  - 1st stage: $E(X) = \beta_0 + \beta_1 G$ ; 2nd stage: $E(Y) = \alpha_0 + \alpha_1 \hat{X}$

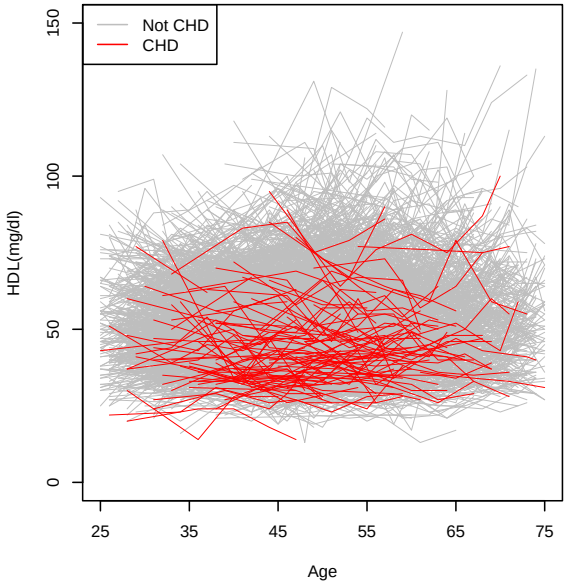
No causal effect of X on Y $\Leftrightarrow H_0: \alpha_1 = 0$

# Mendelian randomization (MR) analysis

- MR has been widely applied to investigate the causal effects of risk factors on disease outcomes using observational epidemiology studies
  - CRP ~ cancer (Allin et al, *JNCI* 2010)
  - HDL ~ myocardial infarction (Voight et al, *Lancet* 2012)
  - BMI ~ cardiometabolic traits and events (Homes et al, *AJHG* 2014)
  - LDL ~ aortic valve calcium and incident aortic stenosis (Smith et al, *JAMA* 2014)
  - Alcohol ~ cardiovascular disease (*BMJ* 2014)
  - Vitamin D ~ mortality (*BMJ* 2014)
  - Telomere length ~ cancer (*HMG* 2015)
  - …

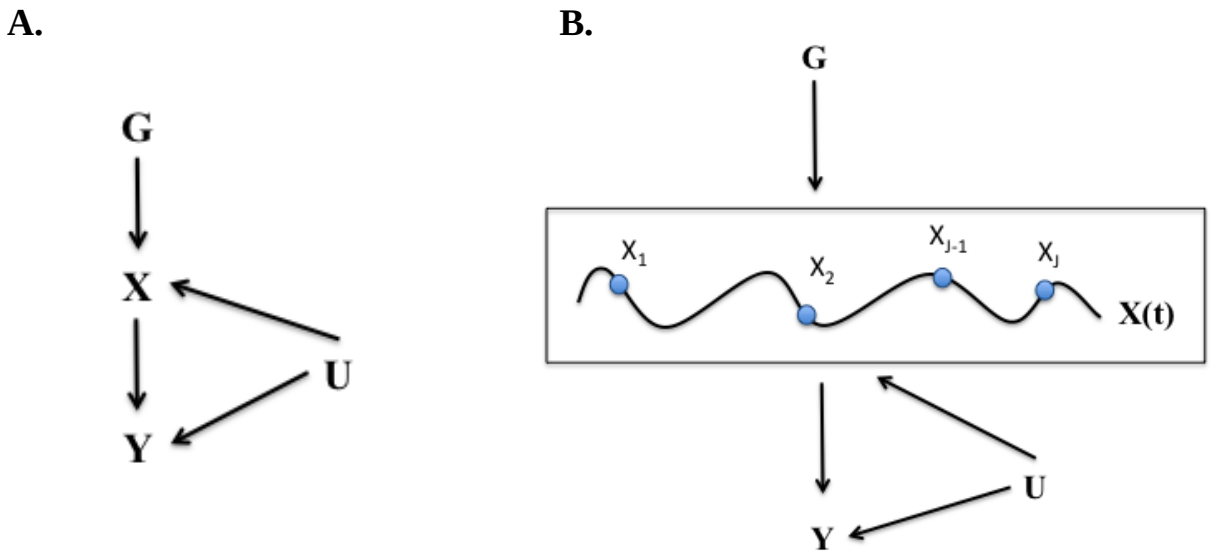


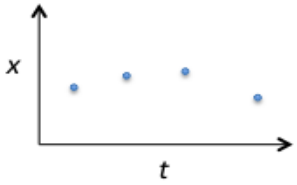




- Framingham Heart Study (FHS) Offspring Cohort
- 7 clinical visits
- Extracted 1709 unrelated individuals
- Exposure of interest: BMI or HDL
- Outcome of interest: incident T2D or CHD

# MR analysis with a time-varying exposure

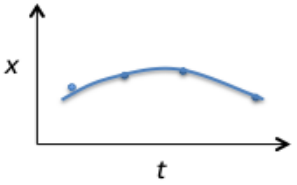





PACE + 2SLS

PACE

PACE + 2SFLR

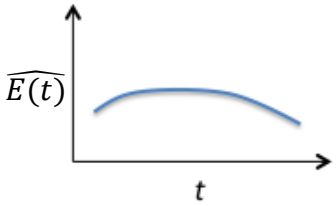



Integration

FLR using IV

$x_i = \int_{T_0}^{T_i} (\hat{x}_i(t) - \hat{\mu}(t))dt$

$x$




$E_i(t) = \beta_0(t) + \beta_1(t)GS_i + v_i(t)$

$x_i = \beta_0 + \beta_1 GS_i + v_i$

LR using IV

Integration

$\hat{x}$

$\hat{x}$

$y_i = \alpha_0 + \alpha_1 \hat{x}_i + w_i$

LR

LR

$y_i = \alpha_0^* + \alpha_1^* \int_{T_0}^{T_i} \hat{E}_i(t)dt + w_i^*.$

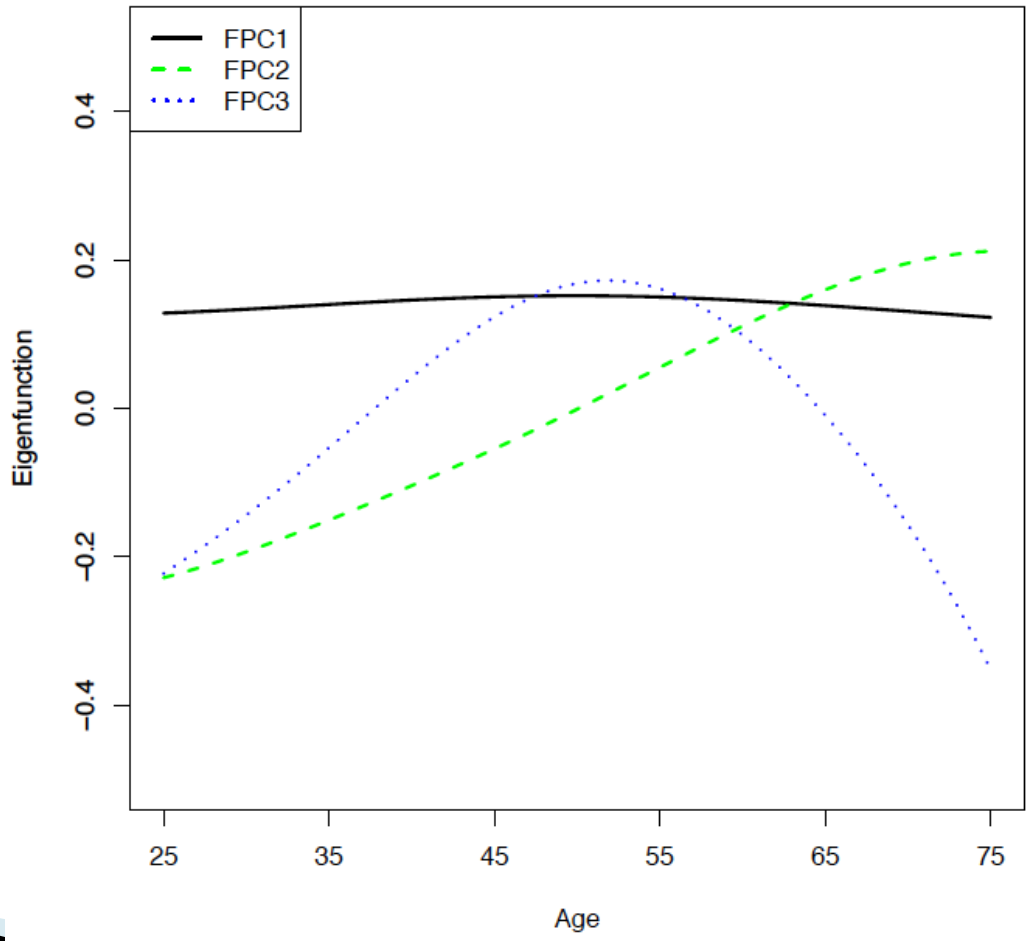
$H_0: \alpha_1 = 0$

Effect of $\hat{x}$ on $y$

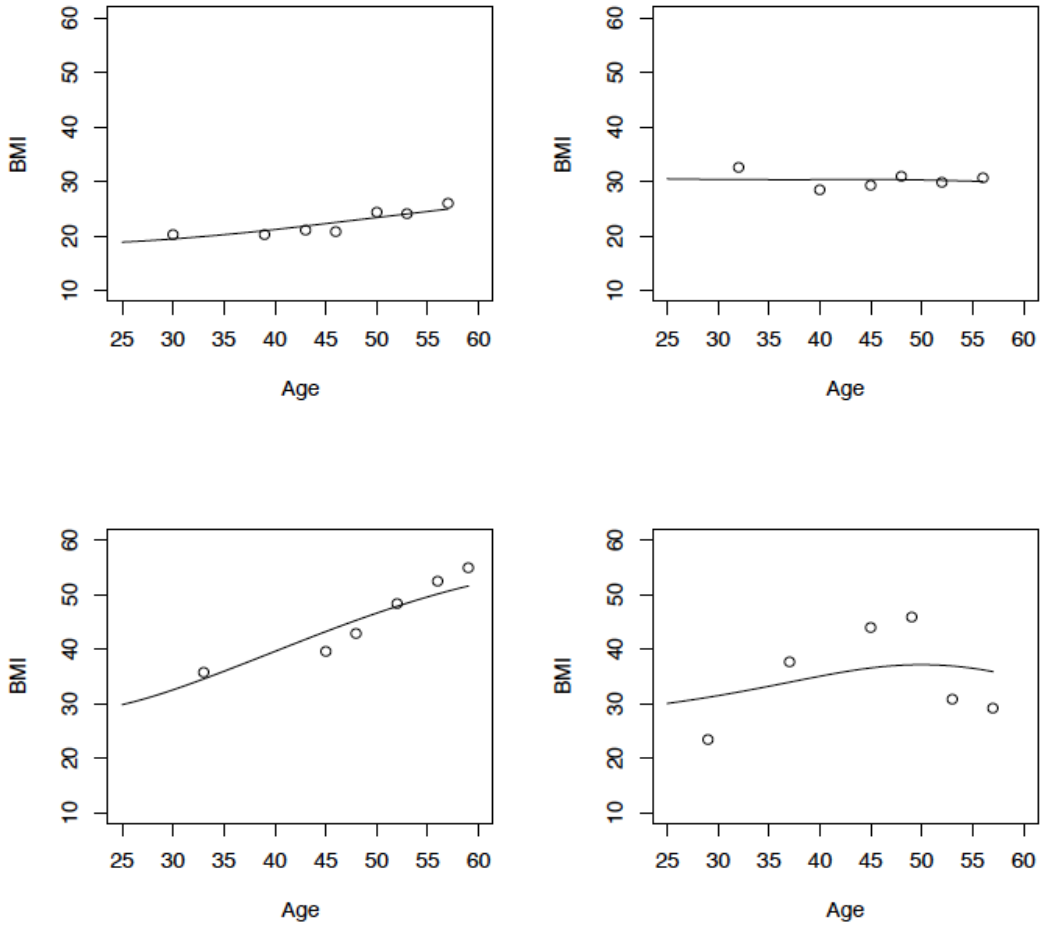
Effect of $\hat{x}$ on $y$

$H_0: \alpha_1^* = 0$

# PACE applied to FHS BMI data



# Observed vs PACE-predicted trajectories



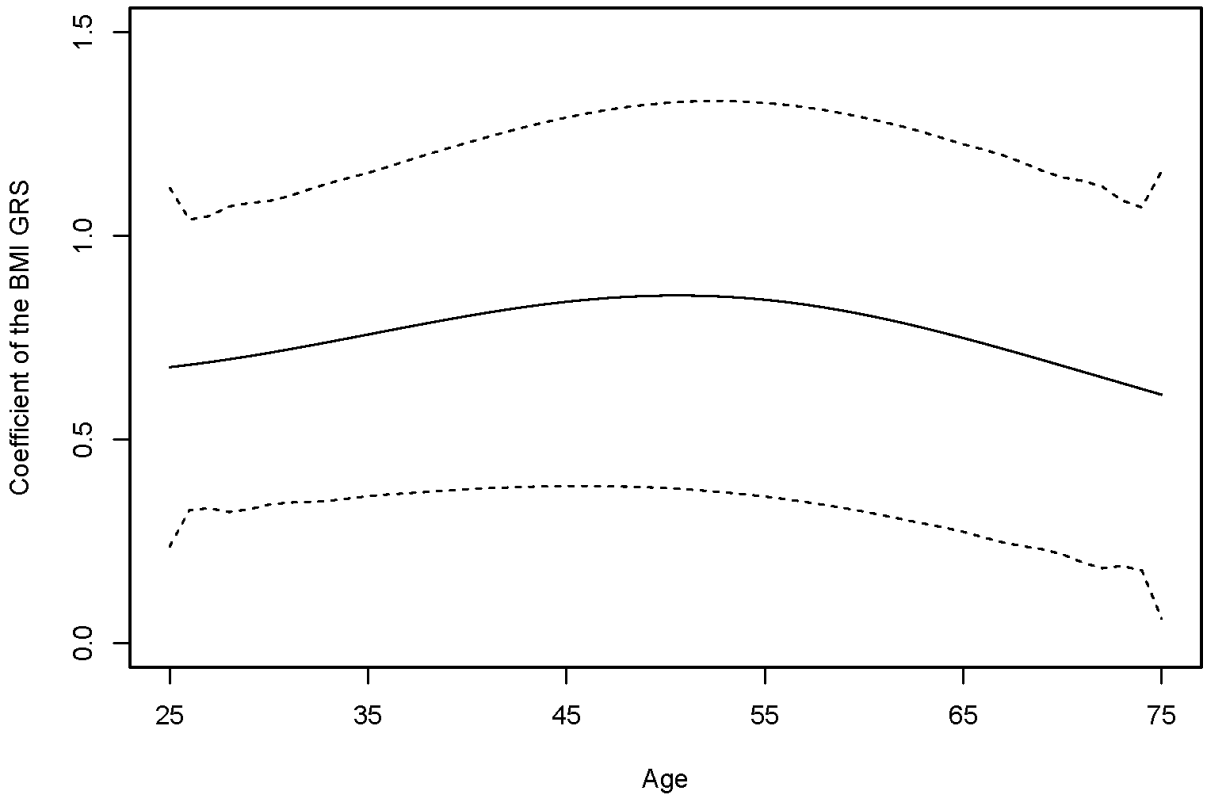
Analysis of the causal effect of BMI on the risk of T2D and CHD, and the effect of HDL on the risk of CHD using the FHS data.

| Exposure | Disease | MR analysis p-value | | | | | Observational analysis p-value | |
|---|---|---|---|---|---|---|---|---|
| | | IV | Baseline 2SRI[a] | PACE +2SRI | PACE+ 2SFRI[b] | IV-outcome test[c] | Baseline | PACE[d] |
| BMI | T2D | GRS | **0.014** | **0.008** | 0.057 | **0.026** | **< 2E-16** | **< 2E-16** |
| | | 14 SNPs | 0.061 | **0.026** | 0.089 | **0.033** | | |
| BMI | CHD | GRS | 0.396 | 0.397 | 0.510 | 0.793 | **0.005** | 0.708 |
| | | 14 SNPs | 0.438 | 0.467 | 0.532 | 0.414 | | |
| HDL | CHD | GRS | 0.518 | 0.531 | 0.239 | 0.967 | 0.114 | **0.005** |
| | | 14 SNPs | 0.783 | 0.790 | 0.489 | 0.083 | | |

[a] 2SRI: 2-stage residual inclusion for binary outcome;
[b] 2SFRI: 2-stage functional residual inclusion;
[c] IV-outcome test: direct association test between IV and disease outcome (VanderWeele et al, Epidemiology 2014);
[d] the cumulative effect of the exposure variable calculated from PACE-recovered curve was tested.

Cao, Rajan, Wei (2015) manuscript under review

# Time-varying effect of the genetic risk score (GRS) on lifetime BMI




$$E_i(t) = \beta_0(t) + \beta_1(t)GS_i + v_i(\mathrm{t})$$

# Analysis of the causal effect of BMI on the risk of T2D by individual clinical visits using the FHS data

| Clinical visit | Sample size | MR analysis p-value | | Observation analysis p-value |
|---|---|---|---|---|
| | | GRS | 14SNPs | |
| 1 | 1515 | **0.023** | 0.167 | 9.58E-22 |
| 2 | 1444 | 0.004 | 0.015 | 6.55E-24 |
| 3 | 1489 | 0.004 | 0.041 | 2.67E-20 |
| 4 | 1591 | **0.017** | 0.012 | 8.37E-22 |
| 5 | 1583 | 0.009 | 0.002 | 1.76E-20 |
| 6 | 1505 | 0.002 | 0.108 | 3.38E-14 |
| 7 | 1440 | **0.097** | 0.390 | 1.75E-08 |
| Bonferroni corrected minP | | 0.014 | 0.014 | 4.59E-23 |

# Part III: Association Analysis of Longitudinal Phenotypes

# Single SNP-based Association Analysis of Longitudinal Phenotypes

- Wei *et al* (2015) Integrative studies of scleroderma-associated genetic and environmental factors with fibroblasts identify polymorphisms of TNFAIP3 in association with MMP expression. *Arthritis & Rheumatology* (in press).
- Systemic Sclerosis (SSc), an autoimmune disease, is a multi-system disorder of connective tissue characterized by extensive cutaneous and visceral fibrosis.
- **Objective**: to demonstrate how human fibroblasts with SSc associated genetic variants respond to time-course and dose-response expression of the extracellular matrix (ECM) genes with silica particle stimulation
- **Samples**: 183 fibroblast strains obtained from skin biopsies of 85 SSc cases and 98 controls
- **Phenotypes**: time-course (24-, 48-, 72-, 96-, 120-hours stimulation with 10 μg silica particles) and dose-response (1, 5, 10, 25 and 50 μg of silica) expression of six ECM genes (*COL1A2, COL3A1, CTGF, MMP1, MMP3* and *TIMP3*)
- **Genotypes**: Illumina Immunochip (~200K genotyped SNPs at 186 autoimmune gene regions and 1000 Genomes-imputed to ~900K SNPs)

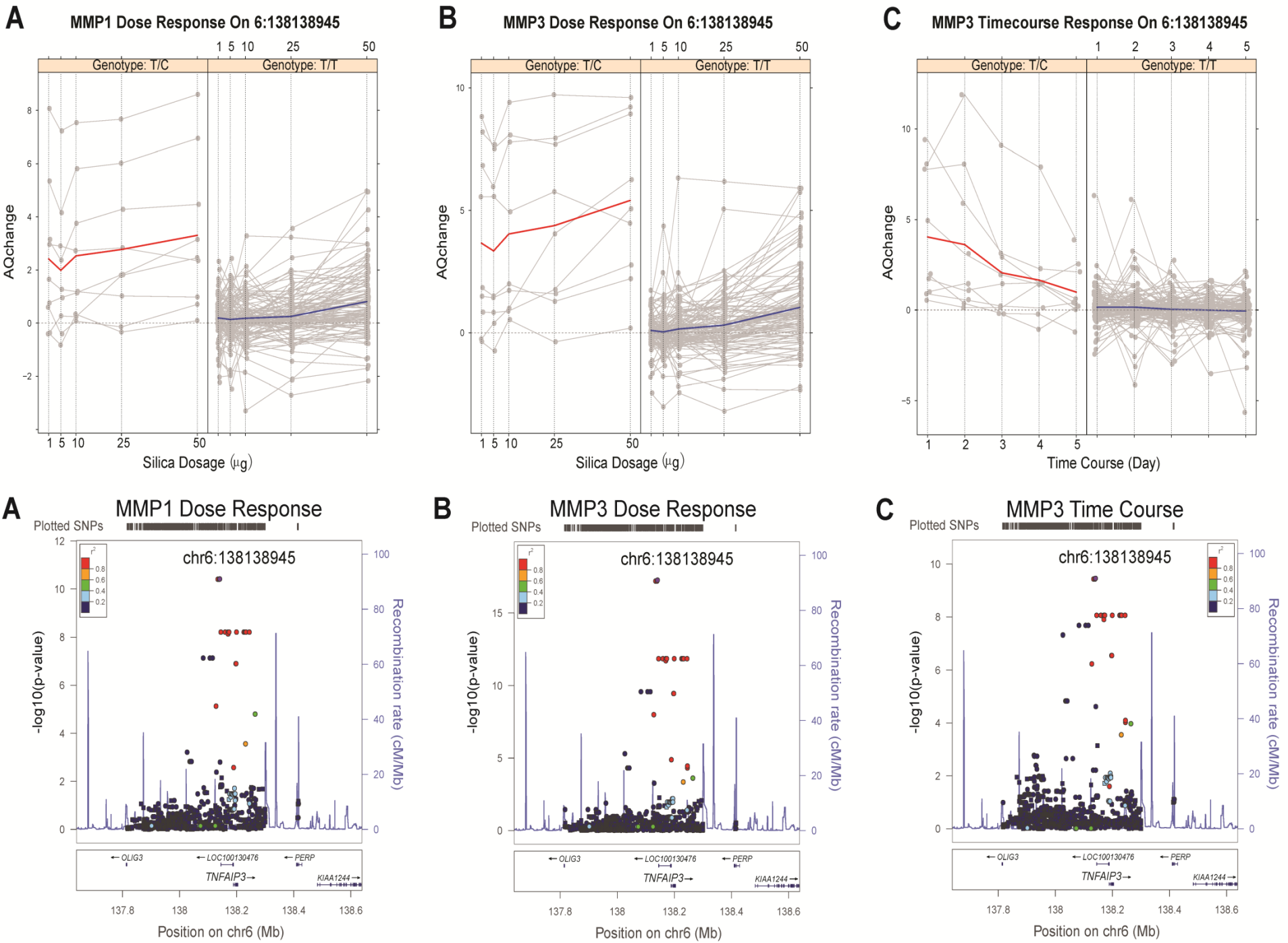
# Single SNP-based Association Analysis of Longitudinal Phenotypes

- Overall analysis strategy: stratified by three races, followed by meta-analysis
  - Non-Hispanic White: 50 cases vs 65 controls
  - African American:13 cases vs 23 controls
  - Hispanics: 22 cases vs 10 controls
- Statistical method: linear mixed model (LMM)

$$y_{ij} = \beta_0 + Z_i\alpha + \beta t_{ij} + \beta_{Gk}G_{ik} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij},$$

  - $y_{ij}$ is a given gene's expression level in subject i at the $j$th measurement ($i = 1,\ldots,$ n and $j = 1,\ldots, 5$);
  - $b_{0i}, b_{1i}$: subject-specific random intercept and slope;
  - $G_{ik}$: # of minor alleles or dosage for SNP $k$;
  - Test $H_0: \beta_{Gk}=0$

Wei *et al* (2015) *Arthritis & Rheumatology*


A
MMP1 Dose Response On 6:138138945
Genotype: T/C
Genotype: T/T
AQchange
Silica Dosage (μg)
B
MMP3 Dose Response On 6:138138945
Genotype: T/C
Genotype: T/T
AQchange
Silica Dosage (μg)
C
MMP3 Timecourse Response On 6:138138945
Genotype: T/C
Genotype: T/T
AQchange
Time Course (Day)
A
MMP1 Dose Response
Plotted SNPs
chr6:138138945
-log10(p-value)
Recombination rate (cM/Mb)
OLIG3
LOC100130476
PERP
TNFAIP3
KIAA1244
Position on chr6 (Mb)
B
MMP3 Dose Response
Plotted SNPs
chr6:138138945
-log10(p-value)
Recombination rate (cM/Mb)
OLIG3
LOC100130476
PERP
TNFAIP3
KIAA1244
Position on chr6 (Mb)
C
MMP3 Time Course
Plotted SNPs
chr6:138138945
-log10(p-value)
Recombination rate (cM/Mb)
OLIG3
LOC100130476
PERP
TNFAIP3
KIAA1244
Position on chr6 (Mb)

# Review: aSPU test for a cross-sectional phenotype

## A Powerful and Adaptive Association Test for Rare Variants

Wei Pan,*,[1] Junghi Kim,* Yiwei Zhang,* Xiaotong Shen,[†] and Peng Wei[1,‡]

Genetics, Vol. 197, 1081–1095 August 2014

- Consider a linear model for a quantitative trait $Y_i$ and a set of *p* rare variants $\boldsymbol{X}_i$,

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i.$$

- To test $H_0: \beta = (\beta_1, \ldots, \beta_k)' = 0$.
- The score vector is $U = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i$ .
- The class of sum of powered score (SPU) test: $T_{SPU(\gamma)} = \sum_{j=1}^{k} U_j^{\gamma}$, for an integer $\gamma \geq 1$.
  - $\gamma = 1 \Longrightarrow$ "burden test";
  - $\gamma = 2 \Longrightarrow$ SSU or SKAT under the linear kernel with equal weighting
  - Increasing $\gamma$ upweights variants with larger $|U_j|$
  - $\gamma = \infty \Longrightarrow$ minimum p-value
- Adaptive SPU (aSPU) test: for $\Gamma = \{1, 2, 3, .., 8, \infty\}$,

$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}$ with p-value obtained by residual permutations.

- Depending on the unknown genetic architecture, a different $\gamma$ may be more powerful than others and aSPU is always close to the winner.
- R package: aSPU

# Extension of aSPU for Longitudinal Phenotypes

## Methods: introduction to notation and formula

Suppose for each subject $i = 1, \ldots, n$, we have $k$ total longitudinal measurements

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})'$$

with $y_{im}$ as a element, $p$ SNPs of interest as a row vector

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

with $x_{ij}$ coded as 0,1 or 2 for the count of the minor allele, and

$$z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})$$

as a row vector for $q$ variates.
Thus, we have:

$$X_i = \begin{pmatrix} x_i \\ x_i \\ \vdots \\ x_i \end{pmatrix}, Z_i = \begin{pmatrix} 1 & z_i \\ 1 & z_i \\ \vdots & \vdots \\ 1 & z_i \end{pmatrix}$$

$X_i$ is a $k \times p$ matrix, and $Z_i$ is a $k \times (q+1)$ matrix.

Yang, ..., Wei (2015a) unpublished data

We then have the GLM equation as,

$$g(\mu_i) = \eta_i = Z_i\varphi + X_i\beta = H_i\theta$$

The consistent and asymptotically normal estimates of $\beta$ and $\varphi$ can be obtained by solving the GEE [LZ86]:

$$U(\varphi, \beta) = \sum_{i=}^{n} U_i(\varphi, \beta) = \sum_{i=1}^{n} (\frac{\partial \mu_i}{\partial \theta'})' V_i^{-1} (Y_i - \mu_i) = 0,$$

with

$$\frac{\partial \mu_i}{\partial \theta'} = \frac{\partial g^{-1}(H_i\theta)}{\partial \theta'}, V_i = \phi A_i^{\frac{1}{2}} R_w A_i^{\frac{1}{2}},$$

and

$$A_i = \begin{bmatrix} v(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & v(\mu_{i2}) & 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & v(\mu_{ik}) \end{bmatrix}$$

$R_w$ is a working correlation matrix, for which the working independence correlation, i.e., $R_w = I$, is often used.

*Quantitative traits*

We use the identity link, i.e. $g(\mu_{im}) = \mu_{im}$ and $v(\mu_{im}) = \phi \times 1 = \phi$. Then we have:

$$U = \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \mu_i)$$

$$\widetilde{\Sigma} = \sum_i (Z_i, X_i)' R_w^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' R_w^{-1} (Z_i, X_i)$$

if the assumption of a common covariance matrices across $Y_i$ for $i$ is valid, e.g. for quantitative continuous traits study , we can adopt a more efficient covariance estimator:

$$\widetilde{\Sigma} = \sum_{i=1}^{n} (Z_i, X_i)' \widehat{\mathrm{var}(Y_i)} (Z_i, X_i) = \sum_{i=1}^{n} (Z_i, X_i)' \left( \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'}{n} \right) (Z_i, X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

which is used by default for its better finite-sample performance

Our goal is to detect whether there is any association between the longitudinal trait and the SNPs via testing on hypothesis

$$H_o : \beta = (\beta_1, \beta_2, \ldots, \beta_p)' = 0$$

We have under the null hypothesis with $g(Y_i) = Z_i\varphi$ to obtain $\varphi$ and predict $\hat{\mu} = g^{-1}(Z\hat{\varphi})$. We hereby have score vector under the null hypothesis, with a working independence model, is

$$U(\hat{\varphi}, 0) = (U'_{.1}, U'_{.2})' = \sum_{i=1}^{n} (U'_{i1}, U'_{i2})'$$

where

$$U_{.1} = \sum_i Z'_i(Y_i - \hat{\mu}_i), U_{.2} = \sum_i X'_i(Y_i - \hat{\mu}_i)$$

As $U$ asymptotically follows a multivariate normal distribution under $H_0$, then the score vector for $\beta$ also has an asymptotic normal distribution:

$$U_{.2} \sim N(0, \Sigma_{.2}), \Sigma_{.2} = \widehat{Cov}(U_{.2}) = V_{22} - V_{21}V_{11}^{-1}V_{12}$$

- Multivariate Score test: $T = U'_{.2}\Sigma_{.2}^{-1}U_{.2}^{-1}$, which follows an asymptotic $\chi^2(p)$ Distribution
  - may suffer from power loss when $p$ is large

# LaSPU: aSPU test for longitudinal phenotypes

- The class of sum of powered longitudinal score (LSPU) test: $T_{LSPU(\gamma)} = \sum_{j=1}^{p} U_{.2,j}^{\gamma}$, for an integer $\gamma \geq 1$.
- Adaptive LSPU (**LaSPU**) test: for $\Gamma = \{1, 2, 3, .., 8, \infty\}$, $T_{LaSPU} = \min_{\gamma \in \Gamma} P_{LSPU(\gamma)}$, with p-value obtained by simulation-based approach for common variants or permutations for rare variants
- Other extensions of the LaSPU test are also proposed, e.g., to combine LaSPU and the multivariate score test

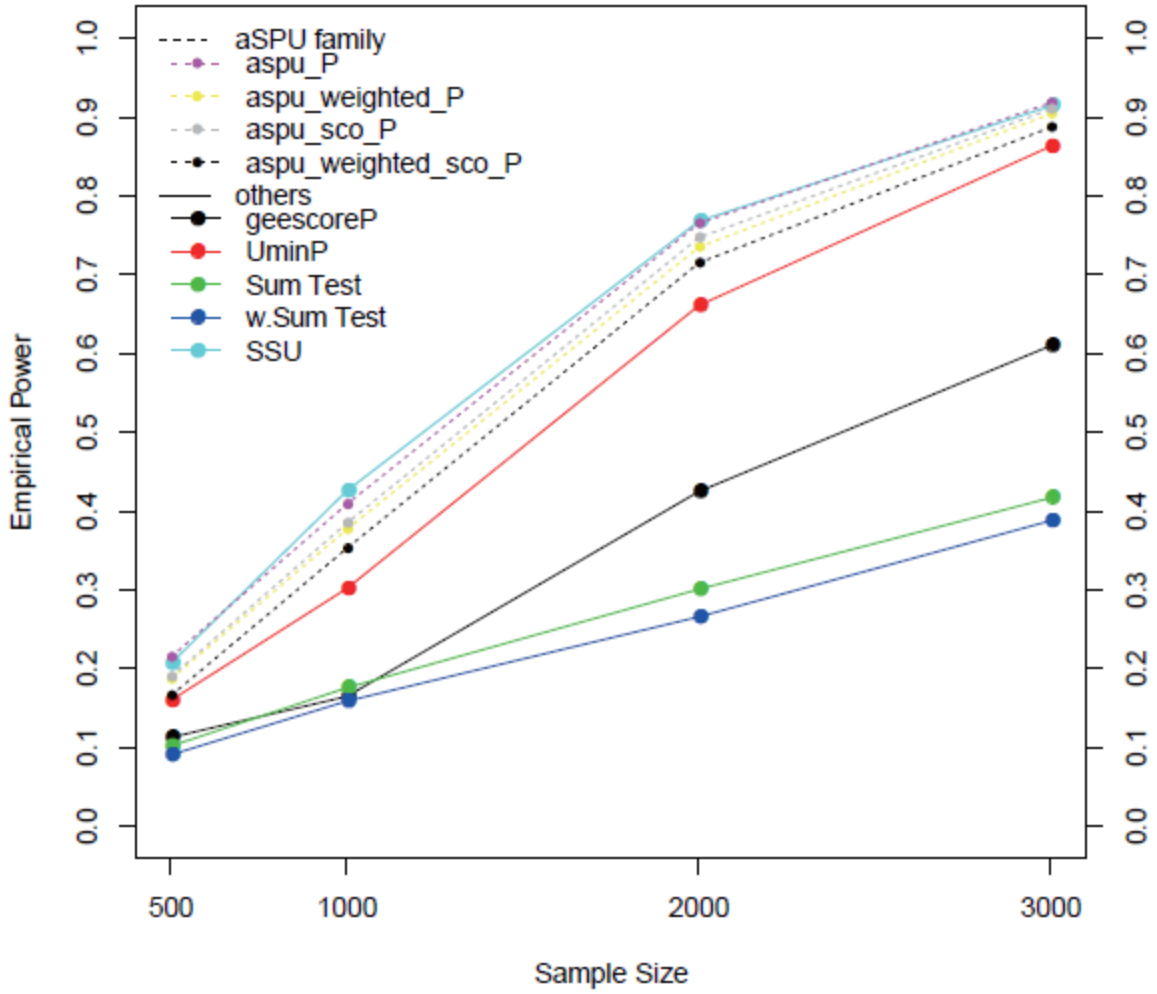
Yang, …, Wei (2015a) unpublished data

# Simulation results: Type I error

| n | pSSU | pSSUw | pScore | pSum | pUminP | SPU(1) | SPUw(1) | SPU(2) | SPUw(2) | LaSPU | LaSPUw | LaSPU.sco | LaSPUw.sco | LaSPU_LaSPUw | LaSPU.omni |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 500 | 0.053 | 0.054 | 0.052 | 0.049 | 0.047 | 0.050 | 0.049 | 0.056 | 0.061 | 0.054 | 0.053 | 0.060 | 0.056 | 0.051 | 0.058 |
| 1000 | 0.055 | 0.040 | 0.042 | 0.048 | 0.054 | 0.048 | 0.049 | 0.056 | 0.043 | 0.047 | 0.045 | 0.052 | 0.051 | 0.046 | 0.051 |
| 2000 | 0.054 | 0.050 | 0.048 | 0.049 | 0.046 | 0.049 | 0.043 | 0.053 | 0.052 | 0.063 | 0.057 | 0.058 | 0.056 | 0.057 | 0.057 |
| 3000 | 0.045 | 0.044 | 0.039 | 0.060 | 0.053 | 0.058 | 0.058 | 0.047 | 0.048 | 0.049 | 0.053 | 0.049 | 0.053 | 0.053 | 0.053 |

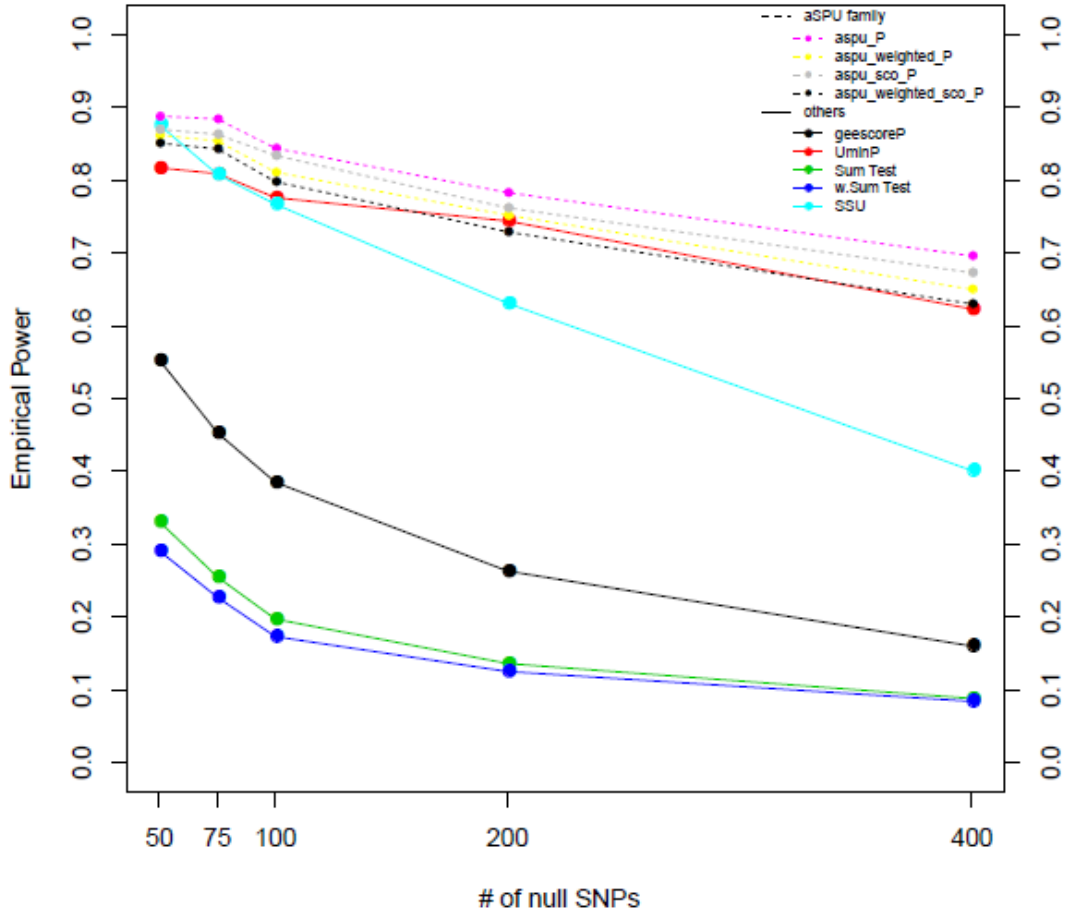
- Type I error rates are well controlled at the nominal level $\alpha = 0.05$ based on 1000 replications

# Simulation results: power



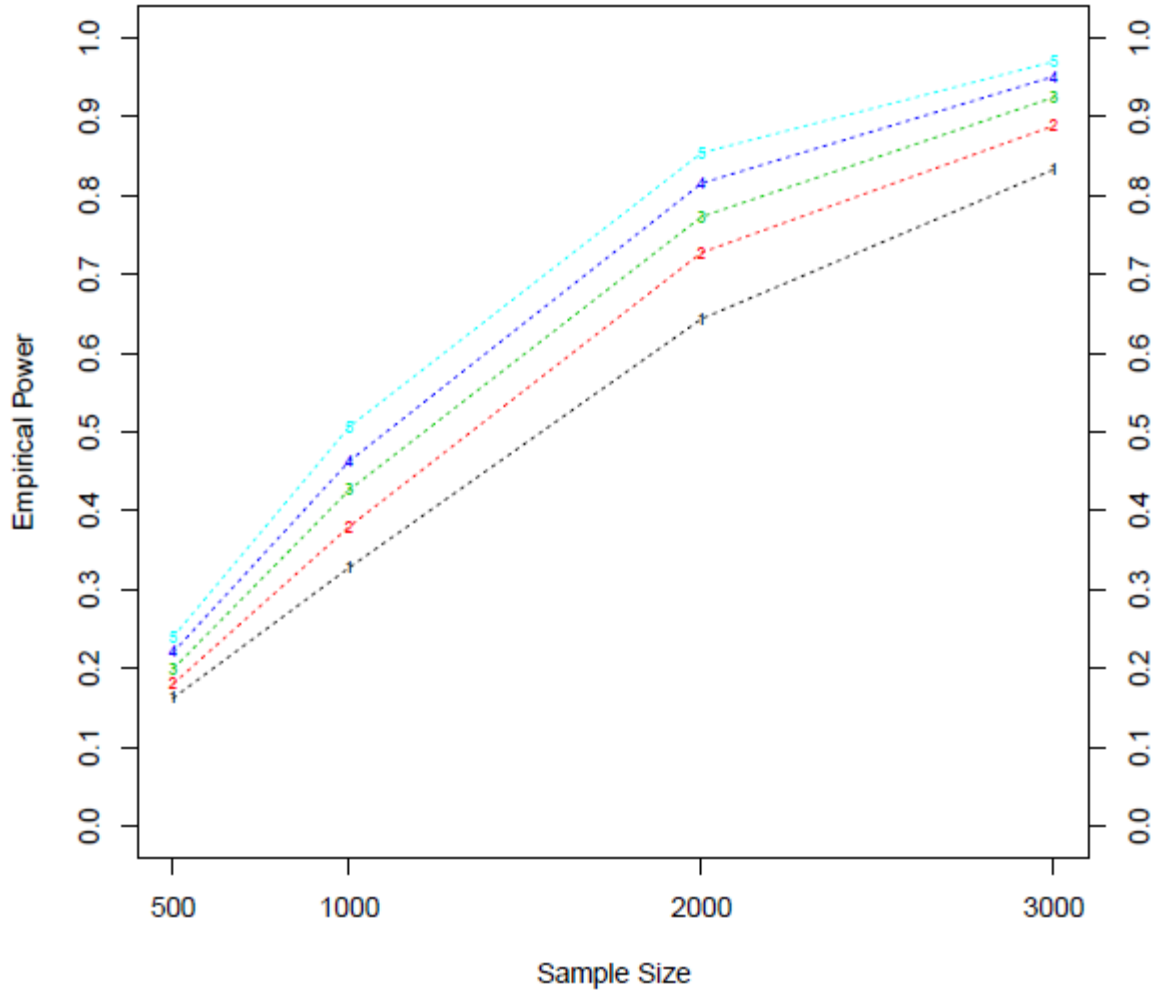

Yang, …, Wei (2015a) unpublished data

# Simulation results: power with increasing number of neutral variants




Yang, …, Wei (2015a) unpublished data

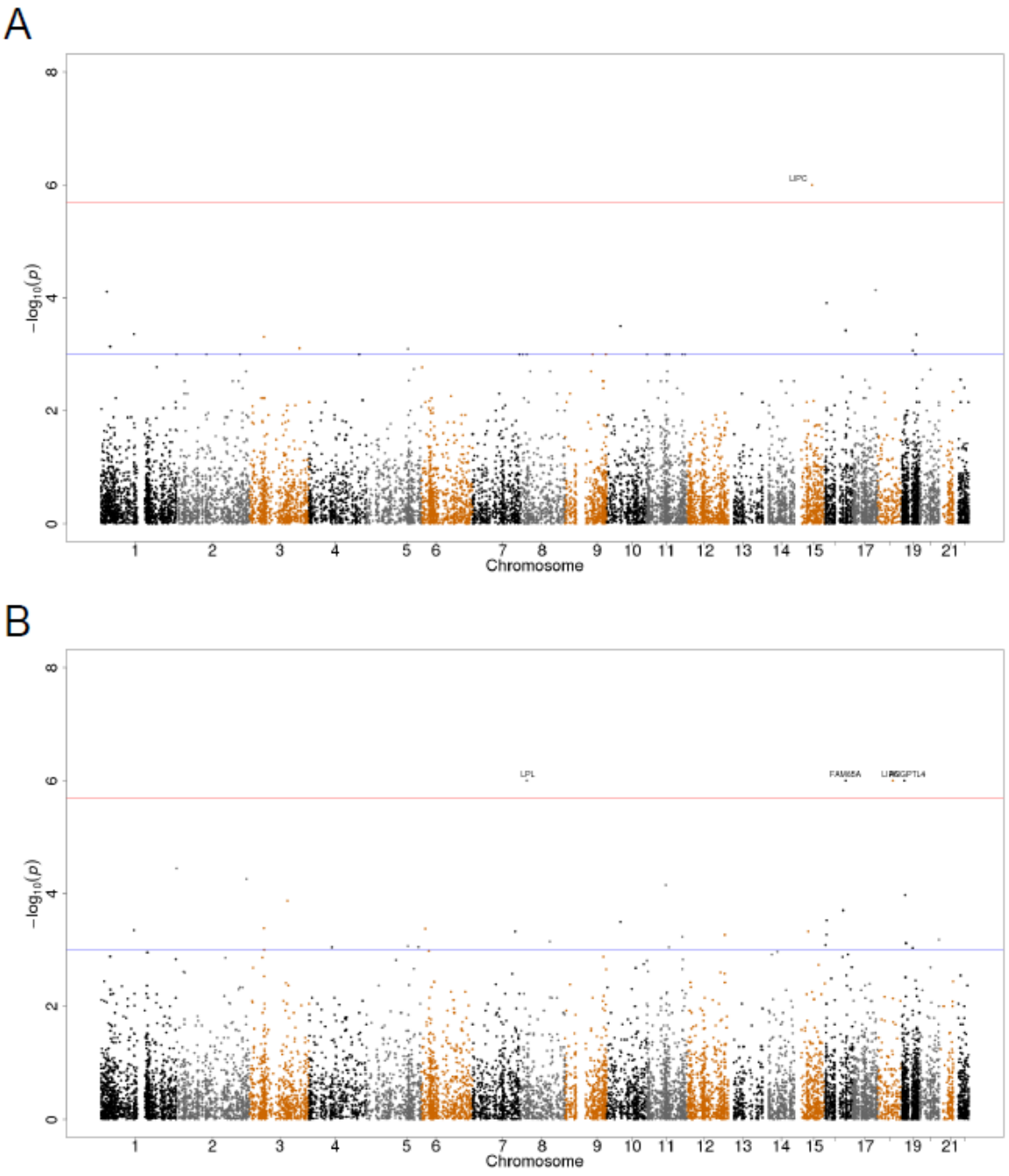
# Simulation results: longitudinal measures increase the power




Yang, …, Wei (2015a) unpublished data

# Application to ARIC ExomeChip Data

- Phenotype: HDL levels measured at four visits in each of n=11,478 ARIC EA subjects
- Genotype: rare/low frequency variants (MAF < 5%) on the ExomeChip
- Statistical methods: LaSPU and its extensions
- Baseline (aSPU) vs all four measures (LaSPU)
- Gene-based association analysis
- Significance threshold at 0.05/20,000=2.5e-06




Baseline HDL-C: *LIPC**

Longitudinal HDL-C: *LPL, LIPG, ANGPTL4* and *FAM65A**

**LIPC* and *FAM65A* were not reported in Peloso et al., AJHG 2014 (n=42,208 with the baseline HDL-C only)

# Top HDL-C associated genes by LaSPU p-values

| Gene | Chr | p Value | No.Variants[a] | CMAC[b] | CMAF[c] | p Value of Baseline[d] |
|---|---|---|---|---|---|---|
| *LPL* | 8 | 1.00E-06 | 10 | 879 | 0.00807 | 9.99E-04 |
| *FAM65A* | 16 | 1.00E-06 | 11 | 751 | 0.00627 | 3.79E-04 |
| *LIPG* | 18 | 1.00E-06 | 11 | 369 | 0.00308 | 3.13E-02 |
| *ANGPTL4* | 19 | 1.00E-06 | 9 | 579 | 0.00591 | 2.89E-01 |
| *ANGPTL8* | 19 | 1.06E-04 | 5 | 64 | 0.00118 | 2.07E-01 |
| *APOC3* | 11 | 5.87E-04 | 3 | 21 | 0.00064 | 9.99E-04 |
| *PAFAH1B2* | 11 | 2.19E-03 | 3 | 287 | 0.00879 | 1.50E-02 |

[a] number of variants contributing to the test
[b] cumulative minor allele count of the variants contributing to the test
[c] cumulative minor allele frequency of the variants contributing to the test
[d] aSPU test using baseline only measurement of HDL-C

Yang, ..., Wei (2015a) unpublished data

# Part IV: Incorporating Biological Information into Genetic Association Analysis

# "aSPUpath" for a cross-sectional phenotype (binary or quantitative)

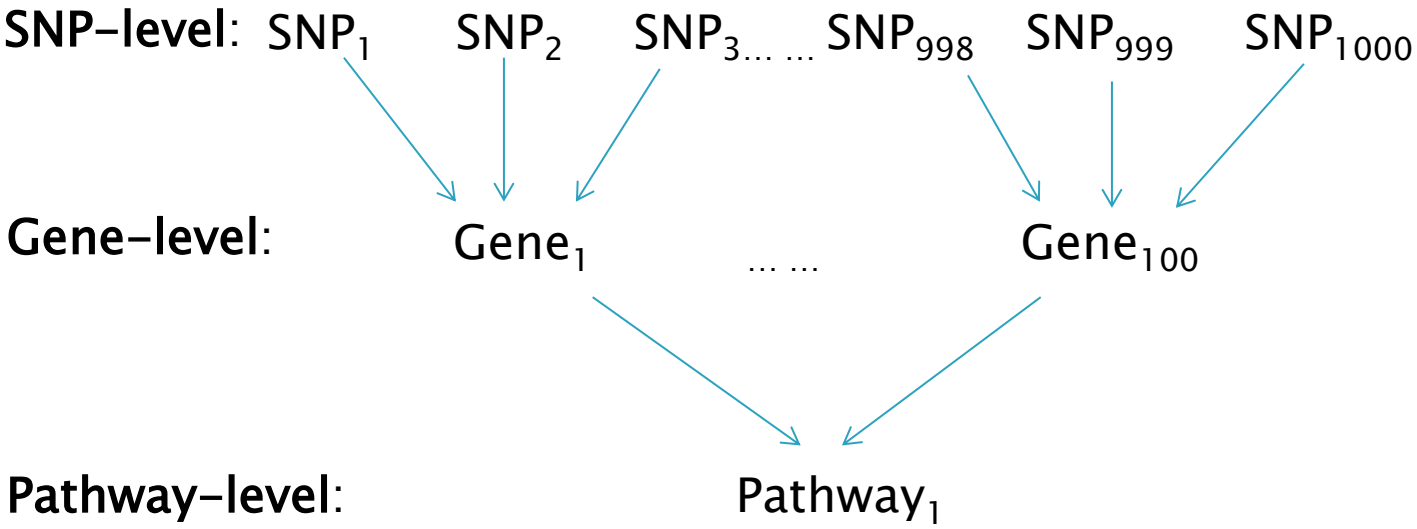
ARTICLE

## A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants

Wei Pan,[1,*] Il-Youp Kwak,[1] and Peng Wei[2,*]

The American Journal of Human Genetics *97*, 86–98, July 2, 2015

Accompanied R package: aSPU



# Power comparison by simulations

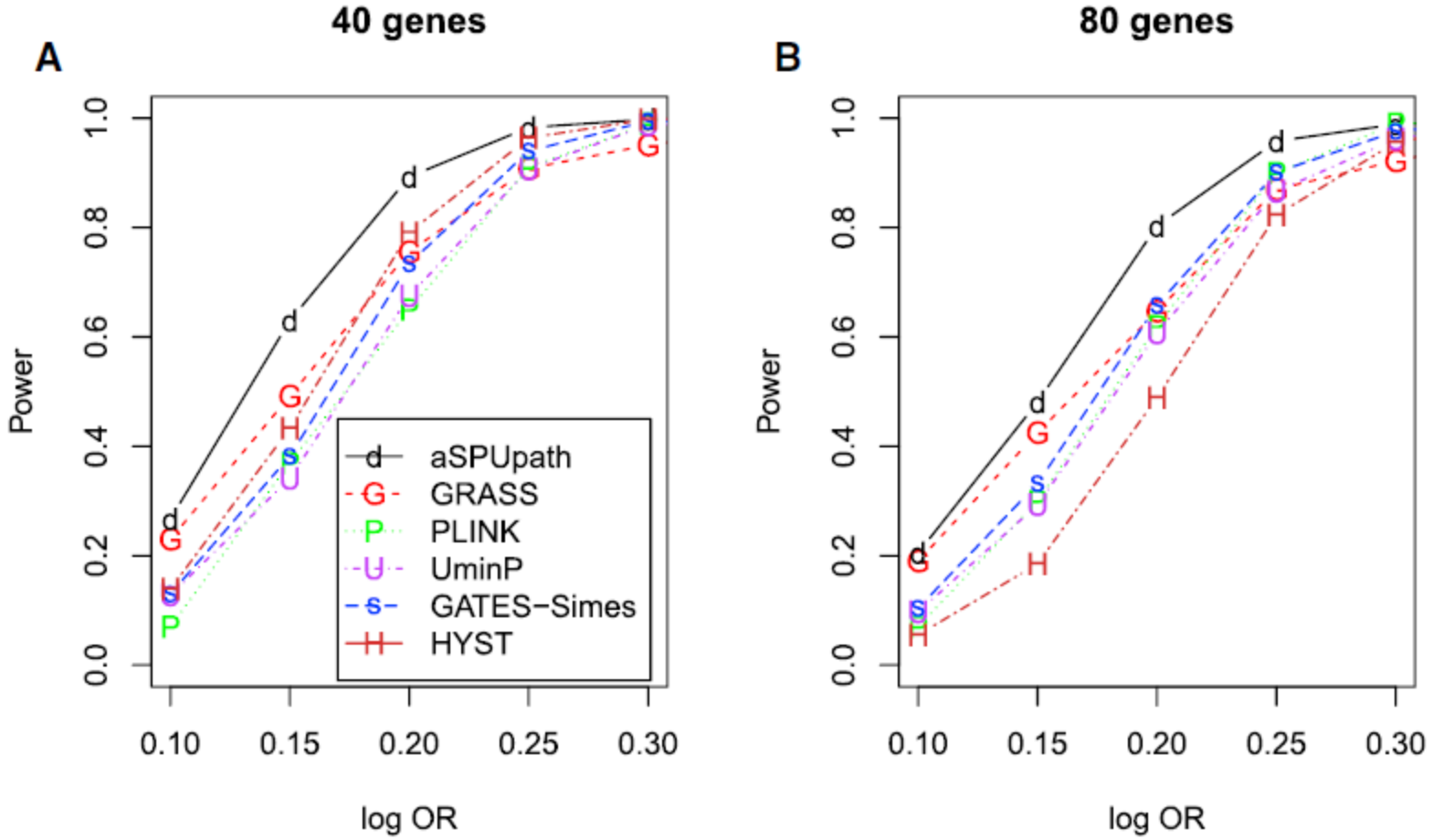
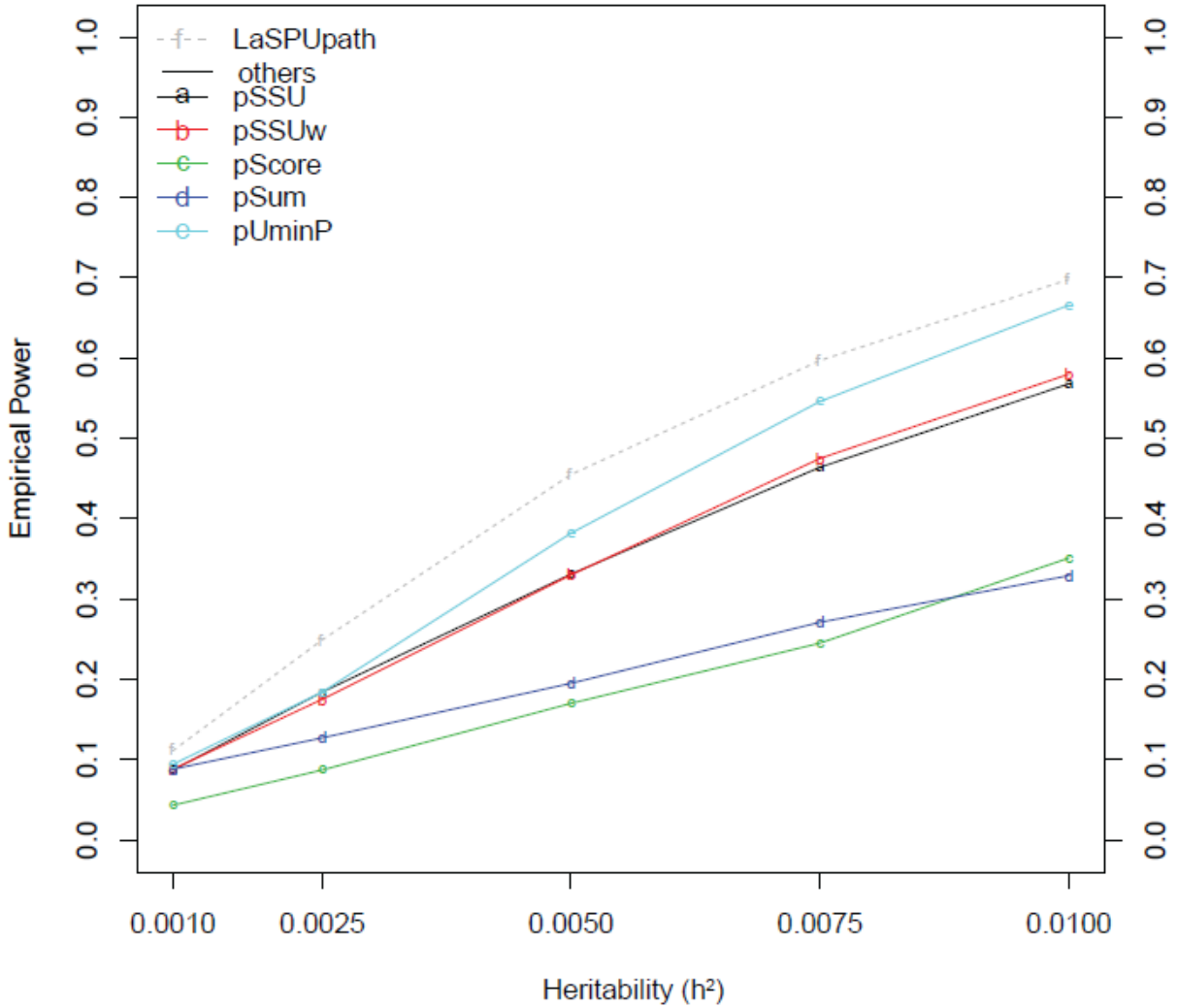


**Table 4. Results of the WTCCC CD GWAS Data Application: KEGG Pathways with p Values < 0.00001 by Any of aSPUpath, GRASS, GATES-Simes, and HYST**

| KEGG ID | Pathway Names | No. of Genes | No. of SNPs | p Values | | | |
|---|---|---|---|---|---|---|---|
| | | | | aSPUpath | GRASS | GATES-Simes | HYST |
| hsa04630 | Jak-STAT signaling pathway* | 145 | 1,410 | <0.00001 | <0.00001 | <0.00001 | <0.00001 |
| hsa04060 | cytokine-cytokine receptor interaction* | 247 | 2,506 | <0.00001 | <0.00001 | <0.00001 | .00001 |
| hsa04660 | T cell receptor signaling pathway* | 105 | 1,373 | <0.00001 | <0.00001 | .00081 | .00021 |
| hsa04310 | Wnt signaling pathway | 143 | 2,087 | <0.00001 | <0.00001 | .00089 | .00238 |
| hsa05310 | asthma | 27 | 271 | <0.00001 | <0.00001 | .00071 | .00002 |
| hsa05330 | allograft rejection | 34 | 466 | <0.00001 | <0.00001 | .00089 | <0.00001 |
| hsa05414 | dilated cardiomyopathy (DCM) | 89 | 2,605 | <0.00001 | <0.00001 | .00382 | .02188 |
| hsa05416 | viral myocarditis | 67 | 1,263 | <0.00001 | <0.00001 | .00148 | <0.00001 |
| hsa04972 | pancreatic secretion | 93 | 2,187 | <0.00001 | .00003 | .00072 | .00211 |
| hsa04621 | NOD-like receptor signaling pathway* | 57 | 502 | <0.00001 | .00542 | <0.00001 | .01012 |
| hsa04062 | chemokine signaling pathway* | 174 | 2,714 | <0.00001 | .00061 | .00131 | .00119 |
| hsa04810 | regulation of actin cytoskeleton | 201 | 3,347 | <0.00001 | .00108 | .00156 | .00962 |
| hsa05131 | shigellosis | 60 | 784 | <0.00001 | .00434 | <0.00001 | .00159 |
| hsa00230 | purine metabolism | 154 | 2,810 | .00759 | <0.00001 | .05376 | .02156 |
| hsa04144 | endocytosis | 180 | 2,575 | .00190 | <0.00001 | .00139 | .01397 |
| hsa04145 | phagosome | 136 | 1,469 | .00101 | <0.00001 | .00314 | .00272 |
| hsa04270 | vascular smooth muscle contraction | 113 | 2,887 | .00025 | <0.00001 | .00086 | .00566 |
| hsa04350 | TGF-beta signaling pathway | 82 | 831 | .00080 | <0.00001 | .00060 | .01381 |
| hsa04514 | cell adhesion molecules (CAMs) | 122 | 3,312 | .00120 | <0.00001 | .00311 | .00043 |
| hsa04612 | antigen processing and presentation | 63 | 543 | .00129 | <0.00001 | .00146 | .00016 |
| hsa04650 | natural killer cell mediated cytotoxicity | 124 | 1,464 | .00199 | <0.00001 | .02586 | .00336 |
| hsa04672 | intestinal immune network for IgA production | 45 | 393 | .00073 | <0.00001 | .00105 | .00009 |
| hsa04940 | type I diabetes mellitus | 39 | 714 | .00031 | <0.00001 | .00102 | <0.00001 |
| hsa05332 | graft-versus-host disease | 33 | 440 | .00036 | <0.00001 | .00086 | .00001 |
| hsa04622 | RIG-I-like receptor signaling pathway | 65 | 474 | .00004 | .0318 | <0.00001 | .30502 |

Asterisks (*) indicate positive control pathways.

# LaSPUpath: extension of aSPUpath to longitudinal phenotypes




Yang, …, Wei (2015b) unpublished data

# LaSPUpath: application to longitudinal HDL-C and ExomeChip data in ARIC EA

- 197 KEGG pathways: each with 10 to 500 genes
- Only rare/low frequency variants: MAF $< 5\%$
- Significance threshold: $0.05/197 = 0.00025$

Table 2: Results of the ARIC Data Application: KEGG Pathways with p Value $< 0.00025$

| KEGG ID | Pathway Name | No. of Genes | No. of SNPs | p Value* | Contributing Genes[a] |
|---|---|---|---|---|---|
| hsa00561 | Glycerolipid metabolism | 44 | 311 | 1.00E-06 | *LPL,LIPG,LIPC, DGKQ,PPAP2A,PNLIPRP3* |
| hsa03320 | PPAR signaling pathway | 55 | 465 | 1.00E-06 | *LPL,ANGPTL4,NR1H3, CD36,APOA1* |
| hsa05010 | Alzheimers disease | 98 | 747 | 1.00E-06 | *LPL,APOE,BID,PLCB3,NDUFS8, NDUFS3,NDUFB6,RYR3,NCSTN* |

[a] p Value of the gene $< 0.05$

*LaSPUpath p-value: based on 1 million permutations

Yang, …, Wei (2015b) unpublished data

# Incorporating ENCODE/RegulomeDB information into association analysis of WGS data

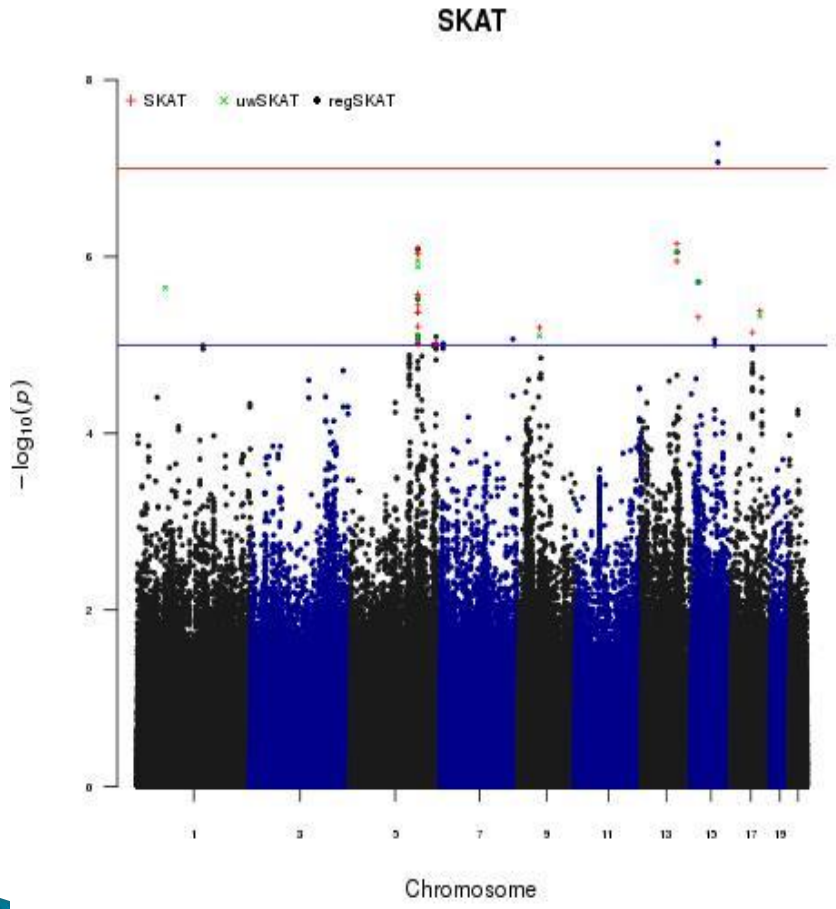
| Score of RegulomeDB | Supporting data |
|---|---|
| 1a | eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak |
| 1b | eQTL + TF binding + any motif + DNase Footprint + DNase peak |
| 1c | eQTL + TF binding + matched TF motif + DNase peak |
| 1d | eQTL + TF binding + any motif + DNase peak |
| 1e | eQTL + TF binding + matched TF motif |
| 1f | eQTL + TF binding / DNase peak |
| 2a | TF binding + matched TF motif + matched DNase Footprint + DNase peak |
| 2b | TF binding + any motif + DNase Footprint + DNase peak |
| 2c | TF binding + matched TF motif + DNase peak |
| 3a | TF binding + any motif + DNase peak |
| 3b | TF binding + matched TF motif |
| 4 | TF binding + DNase peak |
| 5 | TF binding or DNase peak |
| 6 | other |

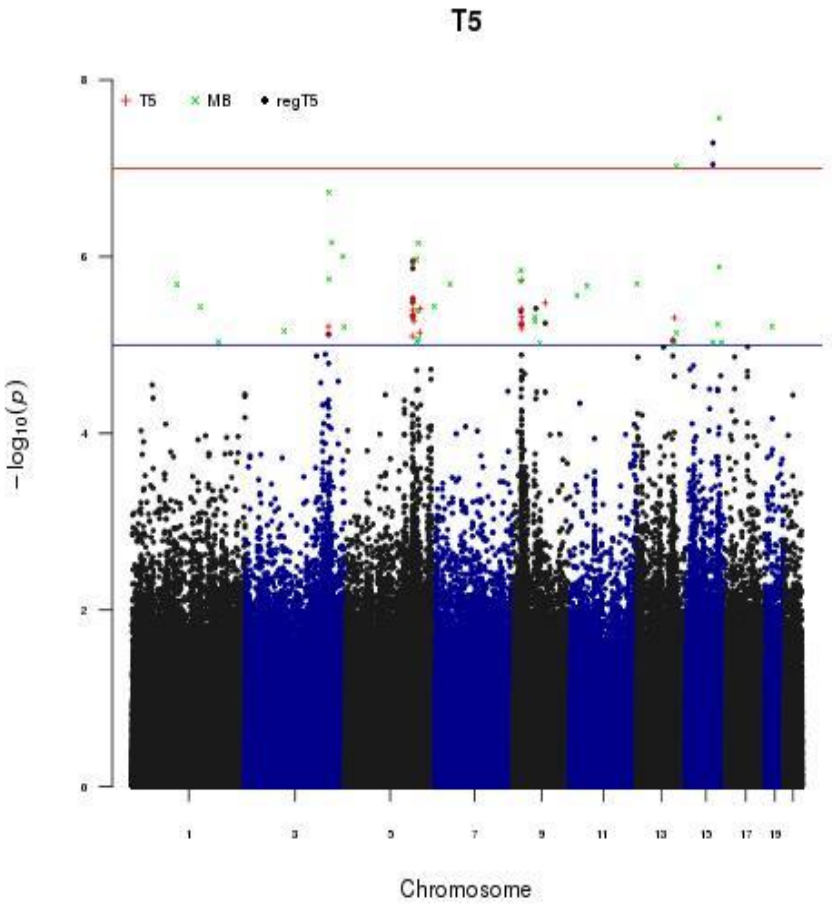
# Analysis of WGS in the Genetic Analysis Workshop (GAW) 19

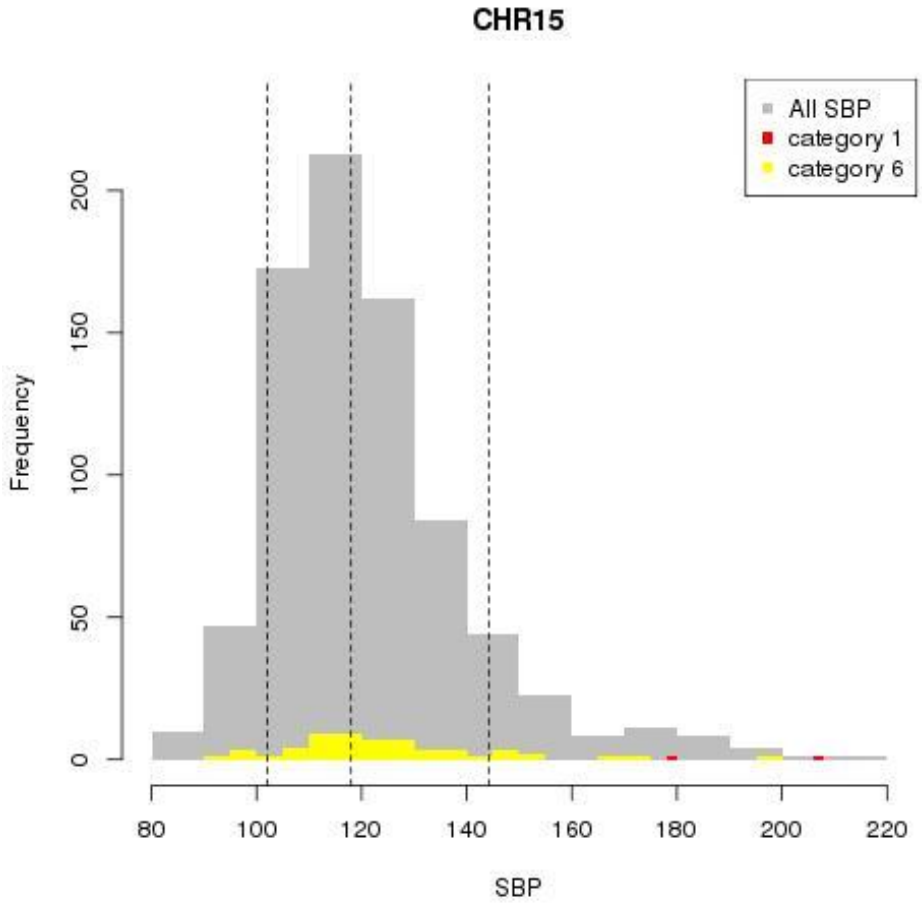
- **Phenotype**: SBP (quantitative trait) of 789 related individuals in the San Antonio Family Studies provided by GAW 19
- **Genotype**: WGS with 8.3 millions variants
  - annotated in RegulomeDB: 0.3%, 2.7%; 2.1%; 7.6%, 30.8%, and 56.6% variants were assigned to category 1, 2, 3, 4, 5 and 6, respectively
- **Statistical methods**: unweighted and weighted T5 and SKAT

Kim and Wei (2015) *BMC Proceedings* (in press)

# Results: Manhattan plots






Distribution of SBP of 789 individuals.

- The sliding window with the center variant chr15:75912109 included variant chr15:75913349 in category 1 (a probably damaging nonsynonymous variant of *SNUPN* by PolyPhen2 with a confidence score of 99.2%)
- The two carriers were half-siblings and had SBP of 179 and 208, respectively) and others in category 6. Dotted lines indicate the 10th, 50th and 90th percentiles of the observed SBP.

# Conclusions

- We have developed and applied new statistical methods for longitudinal data in genetic epidemiology research
  - Gene x Longitudinal exposure interaction
  - Mendelian randomization analysis with a time-varying exposure (Y Cao)
  - Rare variant-based association analysis of longitudinal phenotypes (LaSPU) (Y Yang)
  - Pathway-based association analysis of longitudinal phenotypes (LaSPUpath) (Y Yang)
- Ongoing and future work
  - Rare variant-based GxE tests for baseline and longitudinal phenotypes (T Yang)
  - Biological information-based analysis of whole genome sequencing data (Y Ma)
  - Biological information-based genetic prediction of phenotypes (YM Chen)

# Acknowledgments


- UT School of Public Health
  - Eric Boerwinkle, PhD
  - Alanna C. Morrison, PhD
  - Xiaoming Liu, PhD
- UT MD Anderson Cancer Center, Dept. of GI Medical Oncology
  - Donghui Li, PhD
  - Hongwei Tang, PhD
- Wei Pan, PhD (University of Minnesota)
- My group at UTSPH
  - **Taebeom Kim**, PhD (2014)
  - **Ying Cao**, PhD (2015)
  - **Yang Yang**, PhD candidate
  - Yue-Ming Chen, PhD candidate
  - Tianzhong Yang, PhD student
  - Yiding Ma, PhD student
- Funding support from the NIH: R01CA169122, R01HL116720, and R21HL126032



National Heart Lung and Blood Institute