GENETIC ANALYSES USING FAMILY-BASED SURVEY DATA

Yan Li

Joint Program for Survey Methodolgy

University of Maryland at College Park

yli6@umd.edu

4th Baltic-Nordic Conference on Survey Statistics

Aug 25, 2015

National Genetic Household Surveys (NGHS)

- Conducted in various countries

  e.g. Health 2000 Survey from Finland (Heistaro, 2008);
  Canadian Health Measures Survey (Tremblay et al., 2007);
  U.S. Health and Retirement Study;
  National Health and Nutrition Examination Surveys (NHANES)

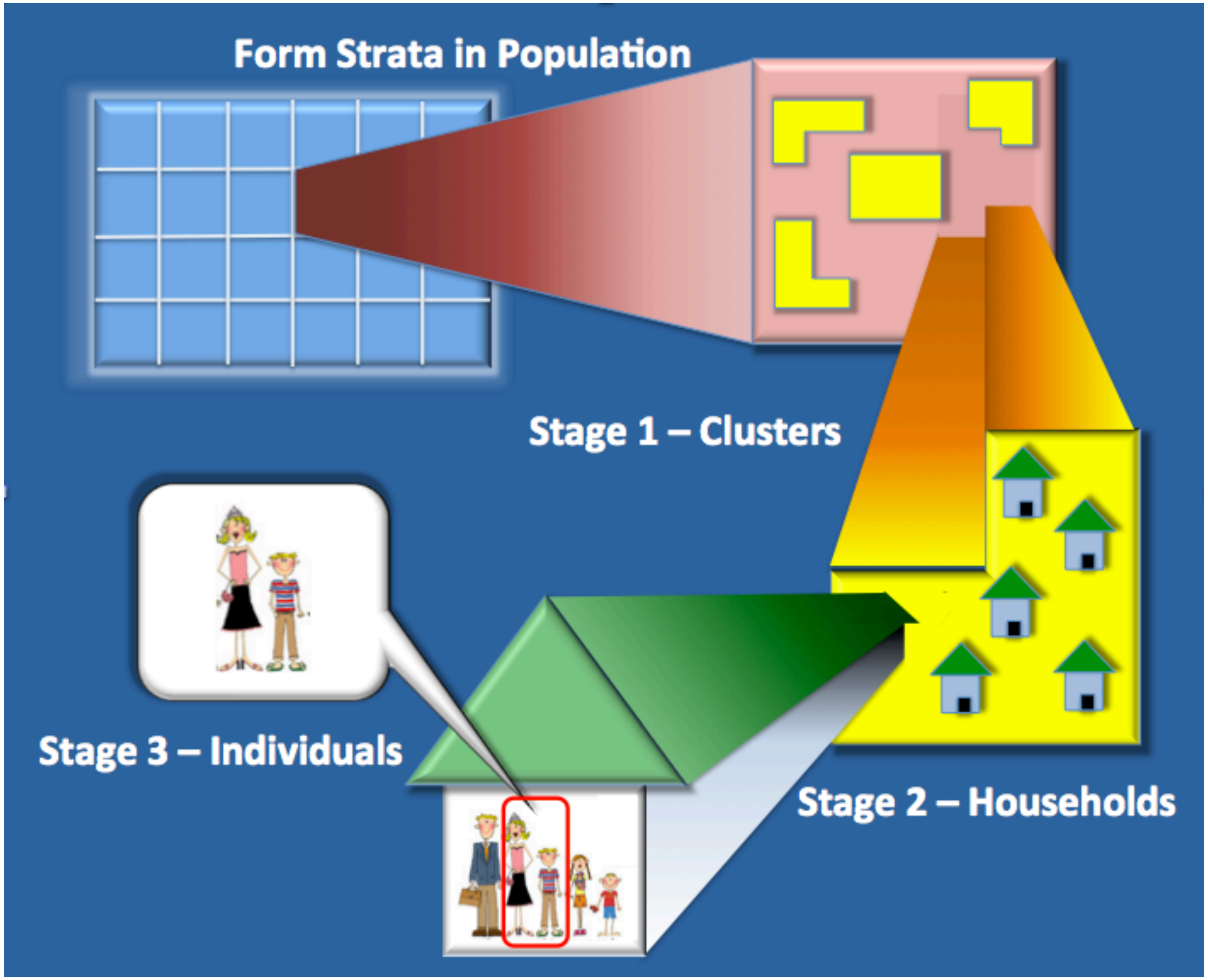
  - Phenotypic, environmental and behavioral data
  - Various types of genetic data

- Less bias in NGHS comparing to traditional genetic studies

  NGHS: random samples representing well-defined populations

  Traditional genetic studies: volunteers or convenience sample

NGHS Cont'd



- Correlation among families due to multistage geographical cluster sampling

- Correlation within families because of biological inheritance

- Differential sampling Weights

OUTLINE

PART I: Hardy-Weinberg Equilibrium Tests

PART II: Genetic Association Studies with Complex Designs

PART I: Hardy-Weinberg Equilibrium Tests

Hardy Weinberg Equilibrium (HWE)

In the case of a single locus with two alleles A and a:

Frequencies of allele A and allle a:$f(A) = p_A$; $f(a) = p_a$

Under ideal conditions,

Hardy Weinberg Equilibrium will be reached after one generation of random mating, i.e., the genotype frequencies remain same:

$f(AA) = p_A^2$; $f(Aa) = 2p_Ap_a$; $f(aa) = p_a^2$

Why Testing HWE is Important?

- Departure from HWE - infer the existence of natural selection, mutation, migration, assertive (non-random) mating, otherwise infer genotyping errors.
- In Genetic Association Studies

  Preliminary step before testing for association between the alleles and disease (Salanti et al., 2005; Zou, 2006; Zou & Donner, 2006)
- HWE is often an assumption in studies testing association of gene-environment interactions with diseases (Chatterjee and Carroll, 2005)

HWE Testing Methods for NGHS

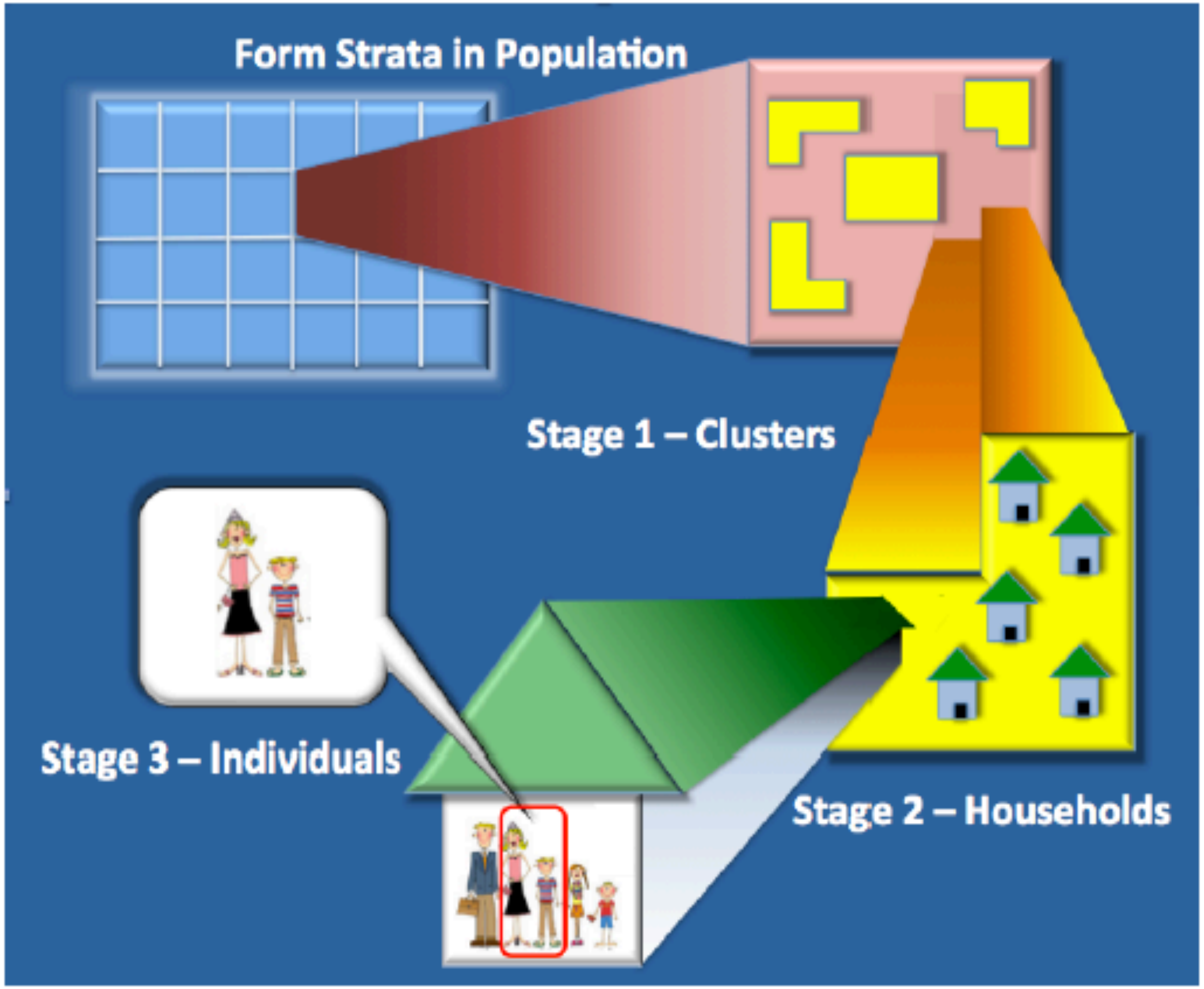
- ➢ Y. Li et al. (2009), Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data. *Biometrics* 65, 1096-104.

  - ✓ **Correlation due to multistage cluster sampling**
  - ✓ **Differential weighting**

How to take account of genetic correlation within families?

METHODS – HWE TESTS

Notations:



H strata

↓

I_h PSUs sampled in the stratum h

↓

J_{hi} families sampled in PSU-hi

↓

K_{hij} individuals in family-hij

For a locus with M alleles (X_1,…, $X_m, \dots, X_M$)

- $p_m = \Pr\{X_m\}$: Frequency of allele X_m
- $p_{mm'} = \Pr\{X_m X_{m'}\}$: Frequency of genotype $X_m X_{m'}$
- $G = \frac{(M+1)M}{2}$: the number of possible distinct genotypes

For example, for a locus with 2 alleles A and a,

M=2

Allele frequencies: p_A and p_a, $p_A + p_a = 1$

Genotype frequencies: p_{AA}, p_{Aa}, p_{aa}, with $p_{AA} + p_{Aa} + p_{aa} = 1$

G=$\frac{(M+1)M}{2} = \frac{(2+1)2}{2} = 3$

- $\boldsymbol{y_{hijk}} = (y_{hijk,1}, \dots, y_{hijk,g}, \dots, y_{hijk,G-1})^T$

  genotype indicators for individual hijk with

$$y_{hijk,g} = \begin{cases} 1 & \text{if the genotype of individual } hijk \text{ is } g \\ 0 & \text{Otherwise} \end{cases}$$

- $\boldsymbol{\mu_{hijk}} = (\mu_{hijk,1}, \dots, \mu_{hijk,g}, \dots, \mu_{hijk,G-1})^T$, where

$$\mu_{hijk,g} = \begin{cases} (1-r)p_l^2 + rp_l & \text{if the genotype } g = \ l/l \\ 2(1-r)p_l p_{l'} & \text{if the genotype } g = \ l/l' \end{cases}$$

- r : **Fixation coefficient** to characterize the departure from HWE – correlation between two alleles in an individual.

Under HWE H$_0$: r = 0

Pseudo Score Function – Individual-based

$$\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{hijk}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{W}_{\boldsymbol{hijk}} \boldsymbol{Var}^{-1}(\boldsymbol{y}_{\boldsymbol{hijk}})(\boldsymbol{y}_{\boldsymbol{hijk}} - \boldsymbol{\mu}_{\boldsymbol{hijk}}(\boldsymbol{\theta})),$$

where

$$\boldsymbol{W}_{\boldsymbol{hijk}} = \begin{bmatrix} \ddots & \cdots & 0 \\ \vdots & w_{hijk} & \vdots \\ 0 & \cdots & \ddots \end{bmatrix}_{(G-1)(G-1)}$$ Inverse of the selection probability

$\boldsymbol{Var}(\boldsymbol{y}_{\boldsymbol{hijk}})$ - covariance matrix of $\boldsymbol{y}_{\boldsymbol{hijk}}$

Working correlation among members within families – Independent

To take account of genetic correlations within families

Pseudo Score Function – Family-based

$$\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{hij}}^{\mathbf{1/2}} \boldsymbol{Var}^{-\mathbf{1}}(\boldsymbol{y}_{\boldsymbol{hij}}) \boldsymbol{w}_{\boldsymbol{hij}}^{\mathbf{1/2}} (\boldsymbol{y}_{\boldsymbol{hij}} - \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})),$$

where

$\boldsymbol{y}_{\boldsymbol{hij}} = (\boldsymbol{y}_{\boldsymbol{hij1}}, \dots, \boldsymbol{y}_{\boldsymbol{hijk}}, \dots, \boldsymbol{y}_{\boldsymbol{hijK_{hij}}})^T$ across selected family members

$\boldsymbol{\mu}_{\boldsymbol{hij}} = E(\boldsymbol{y}_{\boldsymbol{hij}}) = (\boldsymbol{\mu}_{\boldsymbol{hij1}}, \dots, \boldsymbol{\mu}_{\boldsymbol{hijk}}, \dots, \boldsymbol{\mu}_{\boldsymbol{hijK_{hij}}})^T$

Pseudo Estimating Equations

$$\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{hij}}^{\mathbf{1/2}} \boldsymbol{Var}^{-\mathbf{1}}\left(\boldsymbol{y}_{\boldsymbol{hij}}\right) \boldsymbol{w}_{\boldsymbol{hij}}^{\mathbf{1/2}} \left(\boldsymbol{y}_{\boldsymbol{hij}} - \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})\right),$$

where

$\boldsymbol{w}_{\boldsymbol{hij}}$ is sample weight matrix for family-hij with diagonal involving sample weight for each selected family member

$$\boldsymbol{w}_{\boldsymbol{hij}} = \begin{bmatrix} w_{hij1}\boldsymbol{I}_{G-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_{hijK_{hij}}\boldsymbol{I}_{G-1} \end{bmatrix} \text{with}$$

$\boldsymbol{I}_{G-1}$ **=** (G-1) dimensional identity matrix

Pseudo Estimating Equations

$$\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{hij}}^{\boldsymbol{1/2}} \boldsymbol{Var}^{-\boldsymbol{1}}\left(\boldsymbol{y}_{\boldsymbol{hij}}\right) \boldsymbol{w}_{\boldsymbol{hij}}^{\boldsymbol{1/2}} \left(\boldsymbol{y}_{\boldsymbol{hij}} - \boldsymbol{\mu}_{\boldsymbol{hij}}(\boldsymbol{\theta})\right),$$

where $\boldsymbol{Var}\left(\boldsymbol{y}_{\boldsymbol{hij}}\right)$ -- genetic correlation within family-hij

For example, consider family-*hij* with 1 parent (P) and 2 offspring (O_1, O_2) and locus with allele *A* and allele *a*

$$\boldsymbol{Var}(\boldsymbol{y_{hij}}) = \begin{bmatrix} \boldsymbol{\Sigma_P} & \boldsymbol{\Sigma_{P,O_1}} & \boldsymbol{\Sigma_{P,O_2}} \\ & \boldsymbol{\Sigma_{O_1}} & \boldsymbol{\Sigma_{O_1,O_2}} \\ SYS & & \boldsymbol{\Sigma_{O_2}} \end{bmatrix},$$

where

$\boldsymbol{\Sigma_P} = \boldsymbol{\Sigma_{O_1}} = \boldsymbol{\Sigma_{O_2}}$: 2 by 2 covariance matrices between the indicators of genotypes in the same individual

$$\begin{bmatrix} p_A^2(1 - p_A^2) & -p_A^2 \cdot 2p_A p_a \\ SYS & 2p_A p_a \cdot (1 - 2p_A p_a) \end{bmatrix}$$

$\boldsymbol{\Sigma}_{\boldsymbol{P},\boldsymbol{O}_1} = \boldsymbol{\Sigma}_{\boldsymbol{P},\boldsymbol{O}_2}$: covariance between parent and offspring

$$\begin{bmatrix} p_A^3 - p_A^4 & p_A^2 p_a - p_A^2 \cdot 2p_A p_a \\ SYS & p_A^2 p_a + p_a^2 p_A - (2p_A p_a)^2 \end{bmatrix}$$

$\boldsymbol{\Sigma}_{\boldsymbol{O}_1\boldsymbol{O}_2}$: covariance between full siblings

$$\begin{bmatrix} \frac{1}{4}p_A^2 + \frac{1}{2}p_A^3 - \frac{3}{4}p_A^4 & \frac{1}{2}p_A^2 p_a - \frac{3}{2}p_A^3 p_a \\ SYS & p_A p_a - 3p_A^2 p_a^2 \end{bmatrix}$$

$\boldsymbol{\Sigma}$’s are functions of coefficient of condensed identities (CCI), and depend on the family relationship between the pair of individuals

(Lange, 2002 on page 82)

Pseudo Estimating Equations $\boldsymbol{S}(\boldsymbol{\theta}) = \mathbf{0}$:

- Unknown Parameters: $\boldsymbol{\theta} = (\boldsymbol{p}, r)^T$
- $\boldsymbol{S}(\boldsymbol{\theta}) = (\boldsymbol{S}_{\boldsymbol{p}}^T, S_r^T)^T$

Quasi-score test statistic:

$$TS_1 = \hat{S}_r^T\left(\tilde{\boldsymbol{\theta}}\right)\widehat{\boldsymbol{Var}}^{-\mathbf{1}}(\hat{S}_r)\hat{S}_r^T\left(\tilde{\boldsymbol{\theta}}\right),$$

where

$\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{p}}^{\boldsymbol{w}}, r = 0)^T$ – The solution to $\boldsymbol{S}_p\left(\tilde{\boldsymbol{\theta}}\right) = \mathbf{0}$ under H_0

$\widehat{\boldsymbol{Var}}(\hat{S}_r)$ – Consistent estimator of $\boldsymbol{Var}(\hat{S}_r)$

By Taylor linearization method (Rao et al., 1998)

$$\widehat{\boldsymbol{Var}}(\hat{S}_r) = \sum_{h=1}^{H} \frac{I_h}{I_h - 1} \sum_{i=1}^{I_h} \left(\boldsymbol{z}^{hi} - \bar{\boldsymbol{z}}^h\right)\left(\boldsymbol{z}^{hi} - \bar{\boldsymbol{z}}^h\right)^T,$$

where

$$\boldsymbol{z}^{hi} = \sum_{j=1}^{J_{hi}} \left(\frac{\partial \boldsymbol{\mu}_{hij}}{\partial r} - \boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1} \frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{p}} \right) \boldsymbol{w}_{hij}^{1/2} \boldsymbol{Var}^{-1}(\boldsymbol{y}_{hij}) \boldsymbol{w}_{hij}^{1/2} (\boldsymbol{y}_{hij} - \boldsymbol{\mu}_{hij}),$$

$$\boldsymbol{I}_{21} = \frac{\partial}{\partial \boldsymbol{p}} \boldsymbol{S}_r(\boldsymbol{\theta}) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \left\{ - \left(\frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{r}} \right) \boldsymbol{w}_{hij}^{1/2} \boldsymbol{Var}^{-1}(\boldsymbol{y}_{hij}) \boldsymbol{w}_{hij}^{1/2} \left(\frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{p}} \right)^T \right\},$$ and

$$\boldsymbol{I}_{11} = \frac{\partial}{\partial \boldsymbol{p}} \boldsymbol{S}_p(\boldsymbol{\theta}) = \sum_{h=1}^{H} \sum_{i=1}^{I_h} \sum_{j=1}^{J_{hi}} \left\{ - \left(\frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{p}} \right) \boldsymbol{w}_{hij}^{1/2} \boldsymbol{Var}^{-1}(\boldsymbol{y}_{hij}) \boldsymbol{w}_{hij}^{1/2} \left(\frac{\partial \boldsymbol{\mu}_{hij}}{\partial \boldsymbol{p}} \right)^T \right\},$$

evaluated at $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$, and $\bar{\boldsymbol{z}}^h = \frac{1}{I_h} \sum_{i=1}^{I_h} \boldsymbol{z}^{hi}$.

Under suitable conditions (Rao et al 1998),

$$TS_1 = \widehat{\boldsymbol{S}}_{\boldsymbol{r}}^{\boldsymbol{T}}(\widetilde{\boldsymbol{\theta}}) \widehat{\boldsymbol{Var}}^{-\boldsymbol{1}}(\widehat{\boldsymbol{S}}_{\boldsymbol{r}}) \widehat{\boldsymbol{S}}_{\boldsymbol{r}}^{\boldsymbol{T}}(\widetilde{\boldsymbol{\theta}}) \dot{\sim} \boldsymbol{\chi}_{(\boldsymbol{1})}^{\boldsymbol{2}}$$

Simulations show that the developed HWE test TS_1:

- Maintain the nominal level
- Achieve higher power than the test (TS_2) that ignores the genetic correlation within families

Limitations:

Within-family sampling depends on

$$\begin{cases}\text{family relationship (e.g. 1P2O, 3O, etc)} & \surd \\ \textbf{genotype related factors} & \text{X}\end{cases}$$

$$\boldsymbol{w_{hij}^{1/2} Var^{-1}(y_{hij}) w_{hij}^{1/2}}$$

To fix the problem, we use the Pseudo Score Function based on the **pairwise scores**

$$\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}} \boldsymbol{w_{hij}S_{hij}} = \boldsymbol{0},$$

with

$$\boldsymbol{S_{hij}} = \sum_{k=1}^{K_{hij}}\sum_{l=1}^{K_{hij}} \frac{1}{\pi_{kl|hij}} \frac{\partial \boldsymbol{\mu_{hij}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{Var}^{-1}(\boldsymbol{y_{hij}})(\boldsymbol{y_{hij}} - \boldsymbol{\mu}_{hij}),$$

where

○ $\pi_{kl|hij}$ – Joint inclusion probability for pair (*k, l*) given family hij is sampled

○ $\boldsymbol{y_{hij}} = (\boldsymbol{y_{hijk}}, \boldsymbol{y_{hijl}})^{\boldsymbol{T}}$ – a vector of indicators of genotypes for pair of individuals (k, l) in family hij

○ $\boldsymbol{\mu_{hij}} = (\boldsymbol{\mu_{hijk}}, \boldsymbol{\mu_{hijl}})^{\boldsymbol{T}} = \mathrm{E}(\boldsymbol{y_{hij}})$

Quasi-score test statistic (Rao et al. 1998):

~ derived along the same line as above:

$$TS_p = \hat{S}_r^T(\widetilde{\boldsymbol{\theta}})\widehat{\boldsymbol{Var}}^{-\boldsymbol{1}}(\hat{S}_r)\hat{S}_r^T(\widetilde{\boldsymbol{\theta}})\dot{\sim}\chi_1^2$$

Simulations Studies

Population Generation

- 10,000 PSUs with each PSU composed of 40 families
- Generate genotype
  - ○ Consider a biallelic locus (A, a)
  - ○ p_A= p_a= 0.5; r = 0, 0.1, 0.15, 0.2
    - Parents: multinomial distribution with specified genotype frequencies $p(AA) = (1-r)p_A^2 + rp_{A;}\, p(Aa) = 2(1-r)p_A p_{a;}$ $p(aa) = (1-r)p_a^2 + rp_a$
    - Offspring: randomly generated according to Mendelian law
- Population clustering

  Sort all families by #(aa). The 10,000 PSUs are then formed by grouping every 40 families sequentially.

Sampling Designs

- Stage 1: sample 100 PSUs
  - Simple random sampling (srs)
  - Proportional to population size sampling (pps)

    The measure of size related to genotypes, psu's with more #aa is oversampled

- Stage 2: Sample family members - stratified SRS (SSRS) with stratum defined by
  - Family relationship – SSRS(F)
  - Family relationship & genotype – SSRS(GF)

    ~ Oversample genotype *aa*

Test statistics

- TS_1
  - Based on quasi scores at the family level.
  - Considers genetic correlation within families.
- TS_2
  - Based on quasi scores at the family level.
  - Does NOT consider genetic correlation within families.
- TSp
  - Based on quasi pairwise scores within families.

Evaluation Criteria

- RelBias of $\hat{p}_A$

  RelBais ($\hat{p}_A$) = [mean ($\hat{p}_A$) - p_A]/$p_A \times 100\%$

- Variance ratios

  - Analytical variance = Mean of 1,000 estimates of $\widehat{\boldsymbol{Var}}^{\boldsymbol{L}}\widehat{\boldsymbol{S}}_{\boldsymbol{r}}(\widetilde{\boldsymbol{\theta}})$
  - Empirical variance = Variance of 1,000 estimates of p_A

  VR = Analytical variance/Empirical variance

- Rejection Rates at nominal level 5%

  % rejecting r = 0 in 1,000 HWE test

  - Under H0 (r = 0): test size
  - Under H1 (r > 0): power

Figure 1 : Results of 3 HWE tests Under SSRS(F) at 5% nominal level.

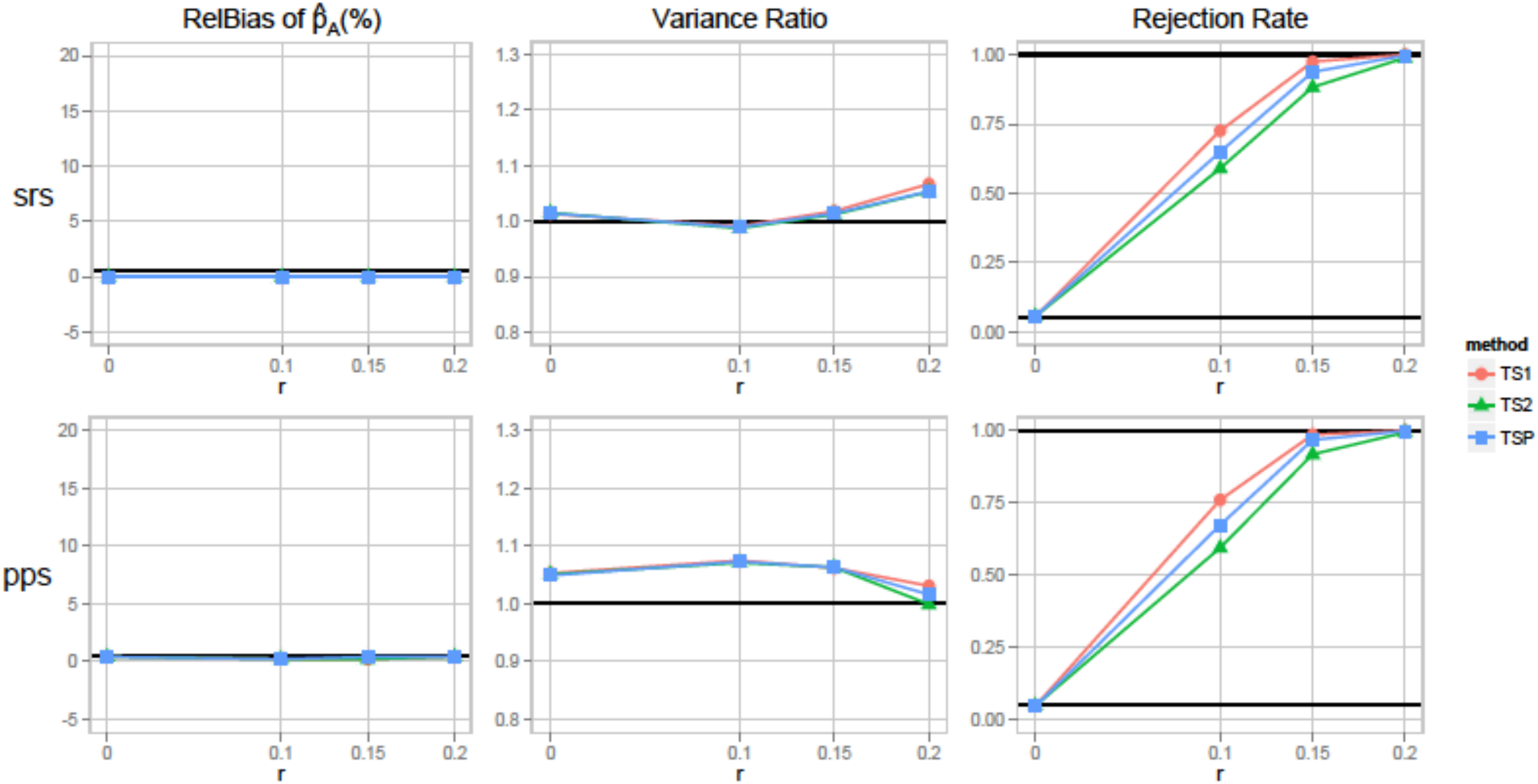
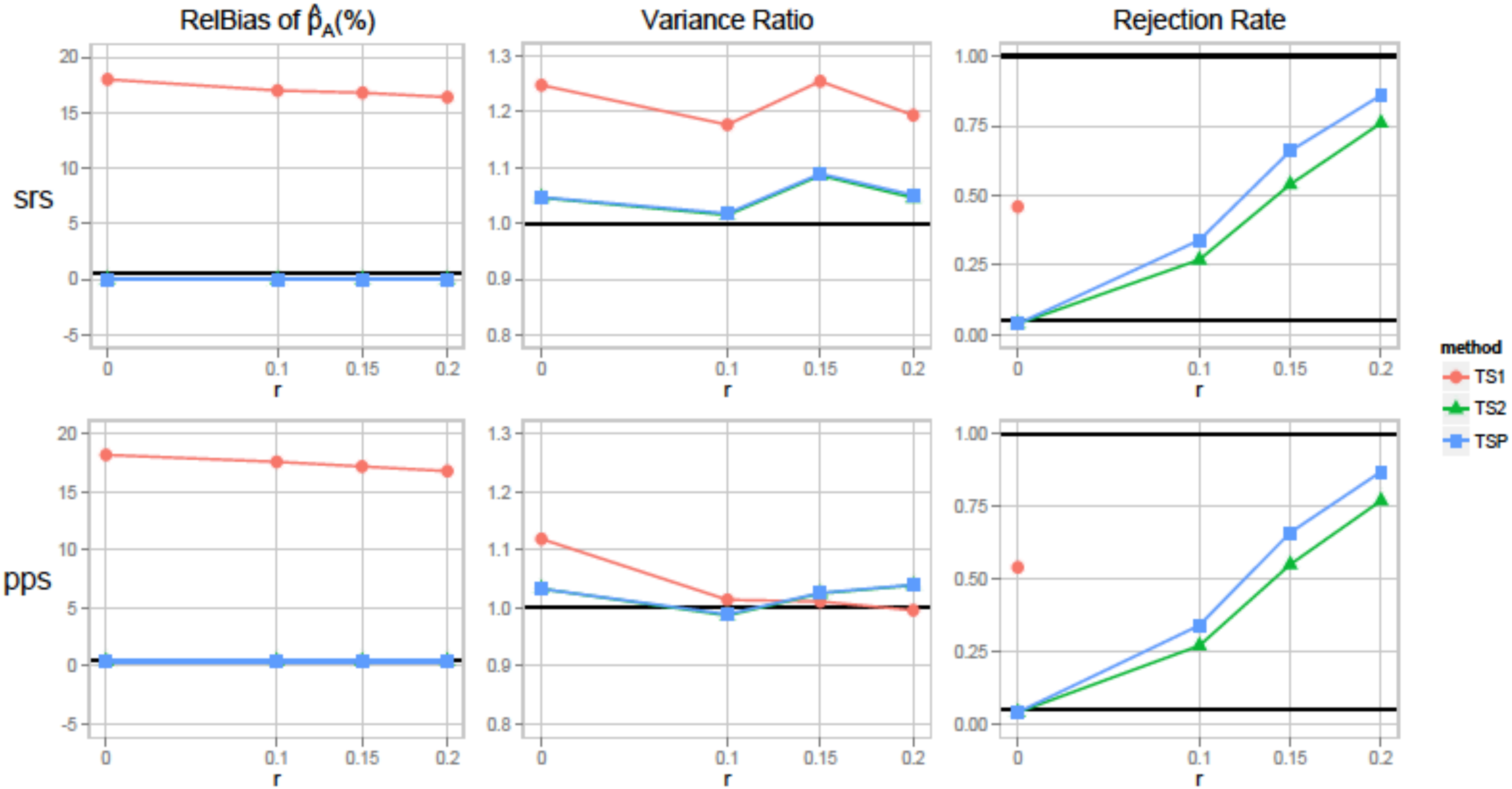


Figure 2 : Results of 3 HWE tests Under SSRS(GF) at 5% nominal level.



IN SUMMARY,

- If within family sampling variables $\perp$ Genotypes

TS_1 prodcues approx. unbiased estimate of allele frequencies, maintains the nominal level at the null hypothesis and achieves the highest power under alternative hypothesis

- If within family sampling variables **related** Genotypes

TS_p prodcues approx. unbiased estimate of allele frequencies, maintains the nominal level at the null hypothesis and achieves the highest power under alternative hypothesis

Conclusions of HWE Tests

- Considers both levels of correlations.
- Considers differential sampling weights

When the within-family sampling is **independent** of genotypes/disease status:

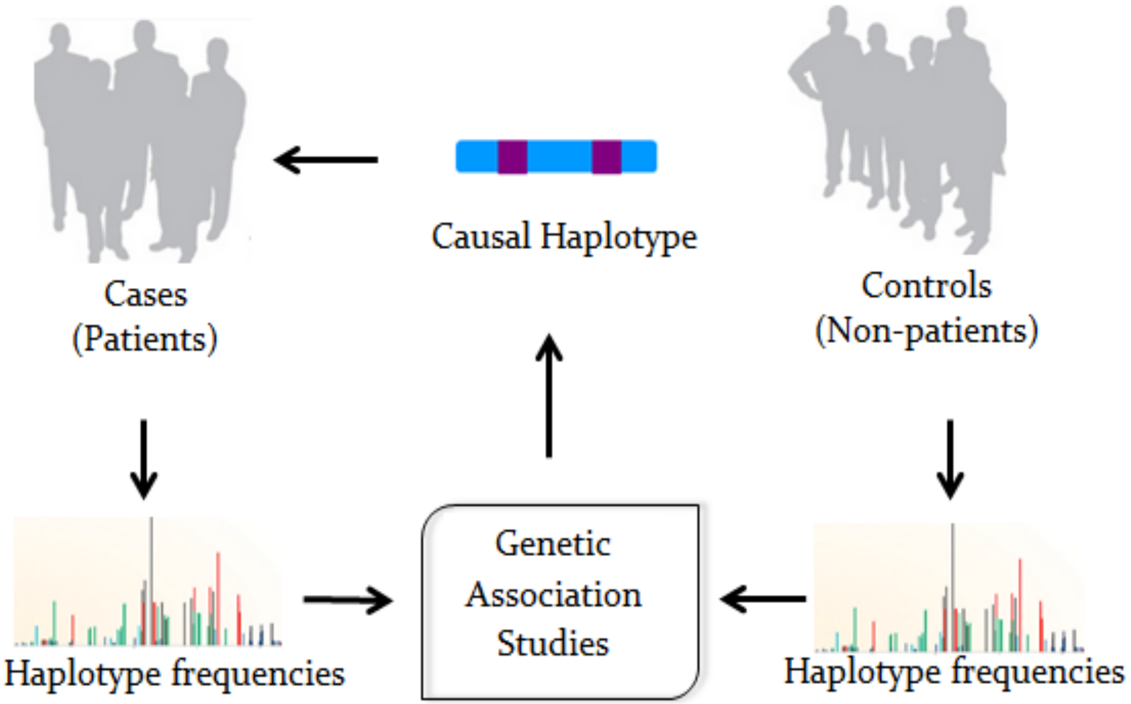
✓Y. Li, et al. (2011), Testing for Hardy Weinberg equilibrium in national household surveys that collect family-based genetic data. *Annuals of Human Genetics* 75, 732-41.

When the within-family sampling is **related** to genotypes/disease status:

✓L. Wang, et al. (2015): A composite likelihood approach in testing for Hardy Weinberg equilibrium using family-based genetic survey data (submitted).

PART II: GENETIC ASSOCIATION STUDIES WITH COMPLEX DESIGN

Genetic Association Studies (GAS) aim to identify genomic variants (e.g., SNPs, haplotypes) that are associated with disease outcomes.



A motivating example—U.S. Kidney Cancer Case-Control Study

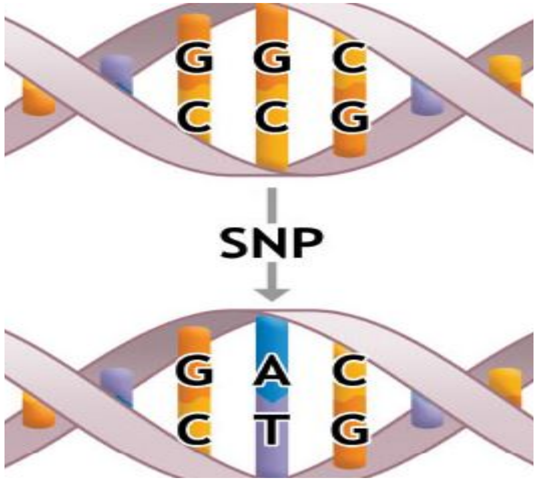
- Population-Based Case-Control Study, **Detroit, Michigan** and Chicago, Illinois
- Cases: identified from the population-based cancer registry in Detroit
- Selection of controls:
  - Stratified Simple Random Sample design
  - Strata defined by the sex, age and black density
- 1,018 cases and 1,038 controls
- Buccal and blood samples were collected as a source of genomic DNA.
- Tobacco use is one of the risk factors of kidney cancer (Brennan et al., 2008)

Analytical Goal 1: Investigate the interaction effect between tobacco use and the SNPs in the *APOE* promoter region (Moore, *et al.* 2009) on the risk of kidney cancer

Analytical Goal 2: Investigate the main effect of the haplotypes inferred from 4 SNPs (Karami et al. 2009) on the risk of kidney cancer.

In GAS, SNP and haplotypes – two common forms of genetic variants

SNP(single-nucleotide-polymorphism) is the occurrence of two or more alleles at one locus in a DNA sequence among individuals in the same population.



www.ibbl.lu

The bases G and A are referred to as alleles,

alternative forms of a DNA segment at a single locus.

Goal 1: Gene-Environment (G-E) Interaction effect on risk of disease

- Standard Logistic Regression Approaches – G-E interaction term included in the regression model (STATA, SUDAAN, R-SURVEY)

  However, Poor power due to small numbers of observations in cells cross-classified genetic variants and exposures.

- *Retrospective* methods can be more efficient – exploring various covariate-distributional assumptions (Chatterjee et al. 2005).

Therefore,

Y. Li and B.I. Graubard (2012), Profile semi-parametric maximum likelihood estimation of gene-environment interaction using population-based case-control study with probability sampling. *Biostatistics*, 13, 711-23.

Analyses results from KCS analysis

	Weighted Logis. Reg.	Pseudo-SPMLE
	Estimates	
Smoking status	0.10	0.30
rs8106922	0.19	0.22
Smoking status×rs8106922	-0.06	-0.19
	Standard Errors	
Smoking status	0.17	0.16
rs8106922	0.13	0.12
Smoking status×rs8106922	0.16	0.11
	p-values	
Smoking status	0.56	0.06
rs8106922	0.15	0.08
Smoking status×rs8106922	0.73	0.09

Goal 2: Haplotype effect on the risk of disease

Haplotype is a set of closely linked SNPs (combination of SNPs) on the same chromosome within the genomic region of interest.

Diplotype is haplotype pairs on homologous chromosomes.

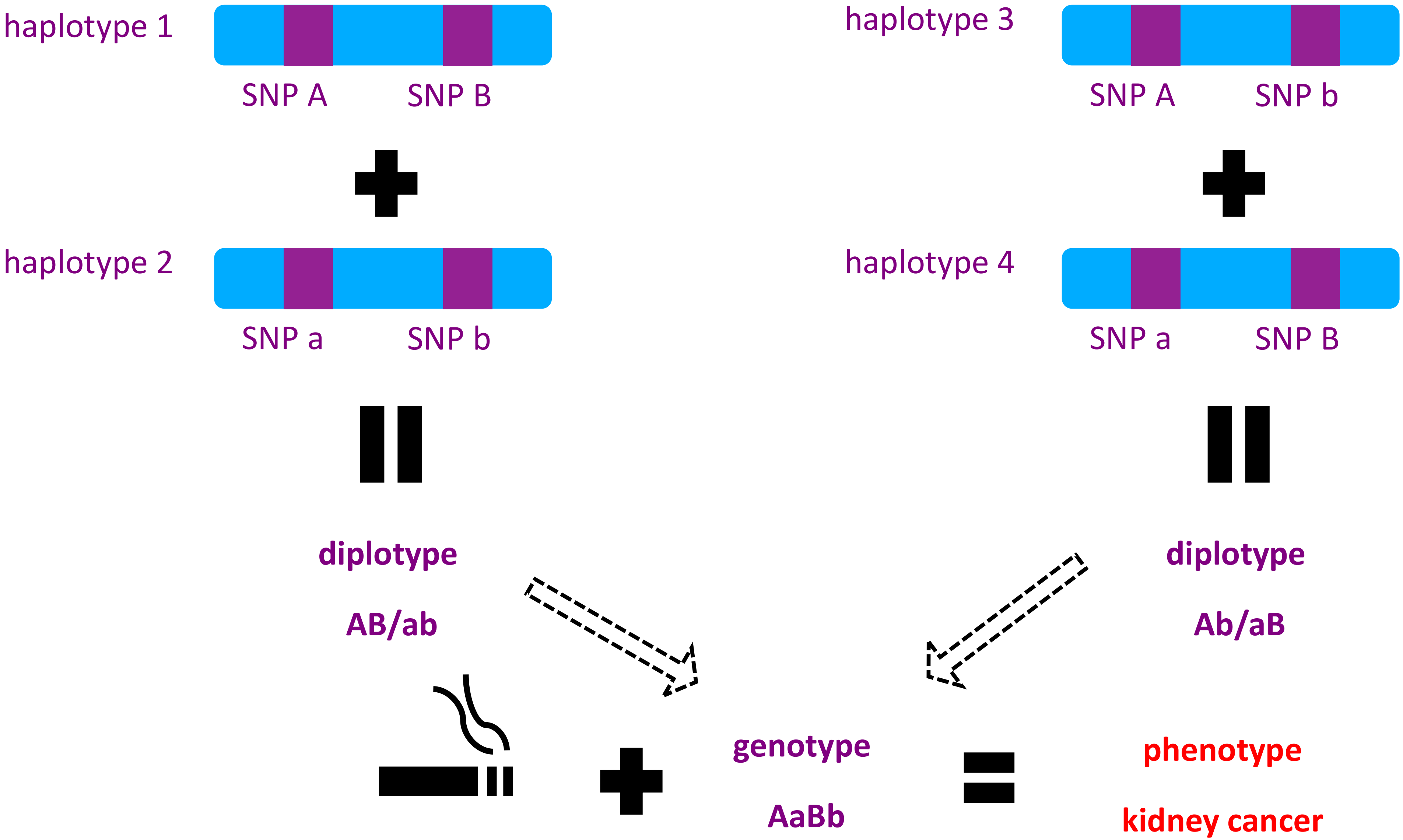
Genotype is a combination of the haplotypes/SNPs on homologous chromosomes.

Phenotype is the traits or conditions that you can observe or diagnose, like eye color or breast cancer.

For a simple example,

Individual 1

Individual 2



haplotype 1

SNP A SNP B

haplotype 2

SNP a SNP b

diplotype

AB/ab

haplotype 3

SNP A SNP b

haplotype 4

SNP a SNP B

diplotype

Ab/aB

genotype

AaBb

phenotype

kidney cancer

Analyzing haplotype data

Advantages

- There is strong evidence that several variants can interact together to have a large effect on the observed phenotype [Schaid, 2004].
- Haplotypes reduce the dimension of association tests and may gain statistical power [Clark, 2004]

Challenges

- Number of haplotypes can be large, and the number is often an unknown priori [Excoffier and Slatkin, 1995].
- **Phase Ambiguity**

  Can genotype data infer which SNPs form the Haplotype?

  NO!

  Phase ambiguity – MISSING DATA PROBLEM

Two-step method

Step 1: Estimation of Haplotype Frequencies $\boldsymbol{\theta}$ – assuming HWE

Challenge: Can be heavy computation if $\boldsymbol{\theta}$ is high dimensional!

Weighted EM algorithm

- ✓ At E-step, the expected number of each haplotype in the population conditional on the genotypes by HWE and
- ✓ At M-step, the weighted estimates of haplotype frequencies,
- ✓ Implemented iteratively until convergence is reached.

The estimate denoted by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{WEM}}$

Step 2: Estimation of Regression Coefficients –Treating $\hat{\boldsymbol{\theta}}_{WEM}$ as fixed

The regression parameters $\boldsymbol{\beta}$ can be obtained by maximizing

$$L_{\beta}^{W}(y, G, E) = \sum_{i=1}^{n} w_i \sum_{j=1}^{c_i} \left\{ \log Pr_{\beta}\left(y_i \middle| E_i, D_i^j\right) Pr_{\hat{\boldsymbol{\theta}}_{WEM},\beta}(D_i^j | obs) \right\}$$

conditional on the observed data *obs=(y, G, E),*

$$Pr_{\hat{\boldsymbol{\theta}}_{WEM},\beta}\left(D_i^j \middle| obs\right) = \frac{Pr_{\beta}\left(y_i \middle| E_i, D_i^j\right) Pr_{\hat{\boldsymbol{\theta}}_{WEM}}(D_i^j)}{\sum_{j'=1}^{c_i} Pr_{\beta}\left(y_i \middle| E_i, D_i^{j'}\right) Pr_{\hat{\boldsymbol{\theta}}_{WEM}}(D_i^{j'})}.$$

where

y_i: Binary indicator of presence, *y*=1, or absence, *y*=0, of a disease

E_i: Environmental covariates associated with the *i*th person

G_i: Genotype of the *i*th person

obs=(y, G, E)

D_i^j: The *j*th diplotype that is compatible with genotype G_i

c_i: the total number of diplotypes that is compatible with G_i

$Pr_\theta(D)$: the prior probability of diplotype D

$Pr_\beta(y|E, D)$: the risk of disease given the exposure (*E*) and *D*

$$L_{\beta}^{W}(y,G,E) = \sum_{i=1}^{n} w_i \sum_{j=1}^{c_i} \{\log Pr_{\beta}\left(y_i \middle| E_i, D_i^j\right) Pr_{\hat{\boldsymbol{\theta}}_{WEM},\beta}(D_i^j|obs)\}$$

w_i: Sampling weights

- Cross-sectional studies – Population Weights (PW)
- Case-control studies with rare disease

$\hat{\boldsymbol{\beta}}_{\boldsymbol{WEM}}$- Inefficient due to the large variation of the PWs

→ Rescale the PW of controls [Scott and Wild, 2011]

$\hat{\boldsymbol{\beta}}_{WEM}$ for all the coefficients apart from intercept is design consistent

- **One-step method**

~ Estimate haplotype frequencies $\boldsymbol{\theta}$ and regression parameters jointly $\boldsymbol{\beta}$

○ Construct the pseudo log-likelihood

$$L_{\gamma}^{w}(y,G,E)=\sum_{i=1}^{n} w_i \sum_{j=1}^{c_i}\left\{\log Pr_{\beta}\left(y_i \middle| E_i, D_i^j\right) Pr_{\gamma}(D_i^j | obs)\right\},$$

Unknown parameters $\boldsymbol{\gamma}=(\boldsymbol{\beta},\boldsymbol{\theta})$

○ Solving $\boldsymbol{\gamma}$ directly are tedious and even numerically infeasible

○ Instead of maximizing L^{w} directly – **Extended WEM (EWEM)**

- E-step: Compute the probability of diplotypes given observed data (genotypes, covariates, and outcomes)

$$Pr\left(D_i^j \middle| obs\right) = \frac{Pr_{\hat{\beta}}\left(y_i \middle| E_i, D_i^j\right) Pr_{\hat{\boldsymbol{\theta}}}(D_i^j)}{\sum_{j\prime=1}^{c_i} Pr_{\hat{\beta}}\left(y_i \middle| E_i, D_i^{j\prime}\right) Pr_{\hat{\boldsymbol{\theta}}}(D_i^{j\prime})}.$$

- M-step: maximize the conditional expectation of log-likelihood based on the complete data (i.e. diplotypes, covariates, and outcomes)

$$L_\beta^W(y, G, E) = \sum_{i=1}^{n} w_i \log\left\{\sum_{j=1}^{c_i} \left\{Pr_\beta\left(y_i \middle| E_i, D_i^j\right) Pr(D_i^j | obs)\right\}\right\}$$

- The iteration is continued until convergence criterion is satisfied.

The resulting estimates are denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{EWEM}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{EWEM}}$.

Variance estimation of the pseudo log-likelihood estimators

The pseudo log-likelihood estimators for haplotype frequencies $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are nonlinear functions of the complex sample data.

By Taylor linearization method,

- Variance of one-step estimators $\widehat{\boldsymbol{\beta}}_{EWEM}$, automatically accounting for the variance due to estimating the haplotype frequencies $\boldsymbol{\theta}$.
- Variance of two-step estimators $\widehat{\boldsymbol{\beta}}_{WEM}$, however, ignoring the variance due to estimating the haplotype frequencies $\boldsymbol{\theta}$.

Simulation Studies

- Case-Control Design
- Cross-Sectional Design

Summary of simulation results

- ✓ Under cross-sectional design, the proposed one-step and two-step methods for estimating haplotype frequencies, $\widehat{\boldsymbol{\theta}}_{WEM}$ and $\widehat{\boldsymbol{\theta}}_{EWEM}$, and regression coefficients, $\widehat{\boldsymbol{\beta}}_{WEM}$ and $\widehat{\boldsymbol{\beta}}_{EWEM}$, perform equally well. Note the estimated variances of the one-step estimator $\widehat{\boldsymbol{\beta}}_{EWEM}$ automatically account for the uncertainty of $\widehat{\boldsymbol{\theta}}_{EWEM}$, and therefore are recommended

- ✓ Under case-control design with rare diseases, the two-step estimator $\widehat{\boldsymbol{\theta}}_{WEM}$ with population weights (PW) and $\widehat{\boldsymbol{\beta}}_{WEM}$ with scaled PW are recommended.

U.S. Kidney Cancer Case-Control Study

	Two-Step	Std
	Estimates	
Haplotype 1010	-0.733	-0.427
Smoking Status	-0.128	-0.057
Smoking Status by 1010	0.075	0.006
	Standard Errors	
Haplotype 1010	0.365	0.339
Smoking Status	0.227	0.209
Smoking Status by 1010	0.207	0.199
	p-values	
Haplotype 1010	**0.045**	**0.207**
Smoking Status	0.573	0.783
Smoking Status by 1010	0.717	0.977

Future Work

- ✓ Hardy-Weinberg Equilibrium tests
  - o TS_p test requires ≥ 2 members selected within families; TS_1 test requires within-family selection $\perp$ genotypes
  - o Future work: New HWE test – combining TS_p and TS_1
- ✓ Genetic Association Studies (GAS)
  - o Haplotype-based inference under retrospective framework
  - o Genome Wide Association Studies
  - o Sequencing Data
- ✓ Surveys help improve genetic studies

Complex sampling designs offer unique advantages in GAS

- o Cost- and time-effective;
- o Obtain representative samples;
- o Avoid biased selection of controls and/or cases

Thank you!