A Flexible Likelihood Framework for Mapping Multiple Phenotypes in Sequencing Based Association Studies of Selected Samples

Dajiang J. Liu & Suzanne M. Leal

RICE®

BCM
Baylor College of Medicine

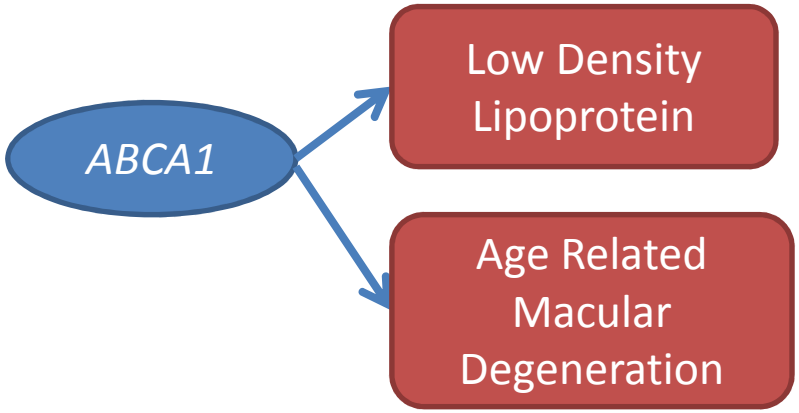
Why Study Pleiotropy and Multiple Phenotypes

Why Study Multiple Phenotypes I

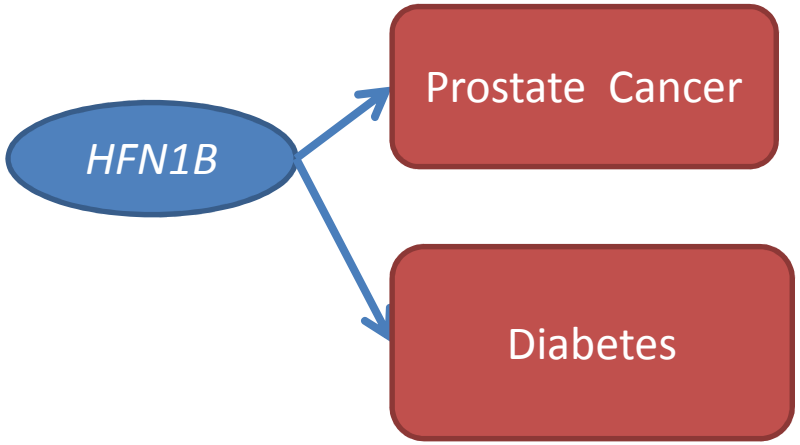
- Pleiotropy: a single gene influences multiple traits
- Gene pleiotropy broadly exists
- Classical example
  - Phenylketonuria (PKU)
    - Mutations in PAH gene cause multiple phenotypes if untreated
      - Mental retardation
      - Reduced hair and skin pigmentation

Why Study Multiple Phenotypes I

- Examples of complex traits from genome-wide association studies (GWAS)



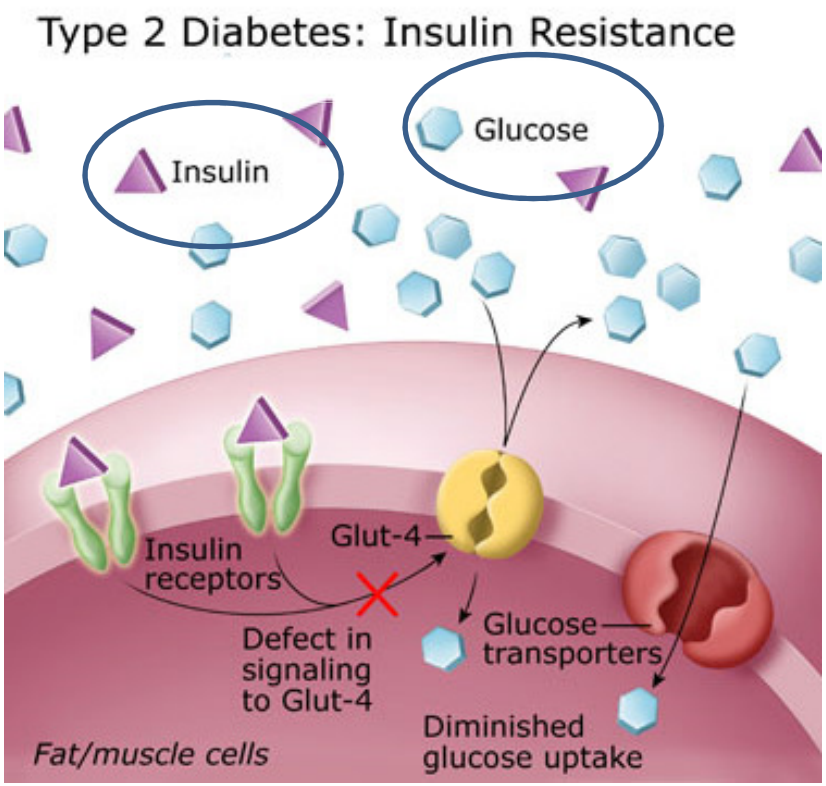
Chen et al *PNAS 2010*



Gudmundsson et al *Nat Genetics 2007*

Why Study Multiple Phenotypes II

- A complex disorder is usually associated with multiple correlated phenotypes:
  - Example: type 2 diabetes
    - Fasting glucose levels
    - Insulin resistance
    - C-reactive protein



Why Study Multiple Phenotypes III

- Studying correlated phenotypes can reveal correlations in the underlying biological pathways

- Mapping multiple phenotypes will
  - Refine phenotypic definitions
  - Reduce sample heterogeneities
    - Example:
      - Etiologies of T2D are hypothesized to be different in obese and non-obese people
      - Stratify samples by body mass index (BMI)

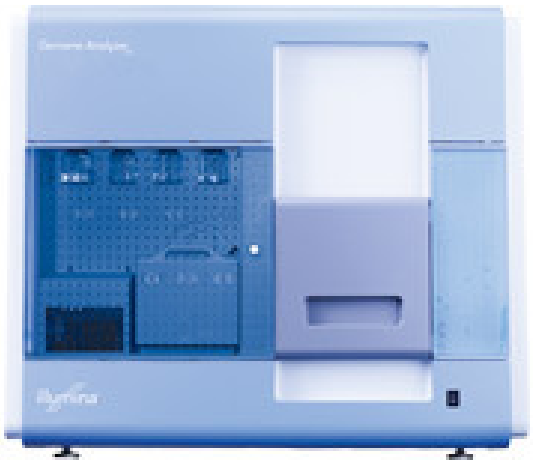
Biological Mechanism for Pleiotropy

- One gene product is used for different biochemical purposes
- One gene product is used in different pathways
- One gene has multiple functional roles
- The gene effect depends on its interactions with other genes

Mapping Secondary Phenotypes in Sequencing Based Genetic Studies

Second Generation Sequencing Platforms

Roche 454



Illumina Solexa

ABI SOLiD

Second Generation Sequencing Technologies

- Much more cost-effective compared to Sanger sequencing
- Already make possible sequencing based genetic association studies
  - When coupled with target enrichment methods, e.g. exon capture
- Still expensive to generate and process sequence data from
  - Large number of individuals
  - At high coverage depth

Sequencing Based Genetic Studies

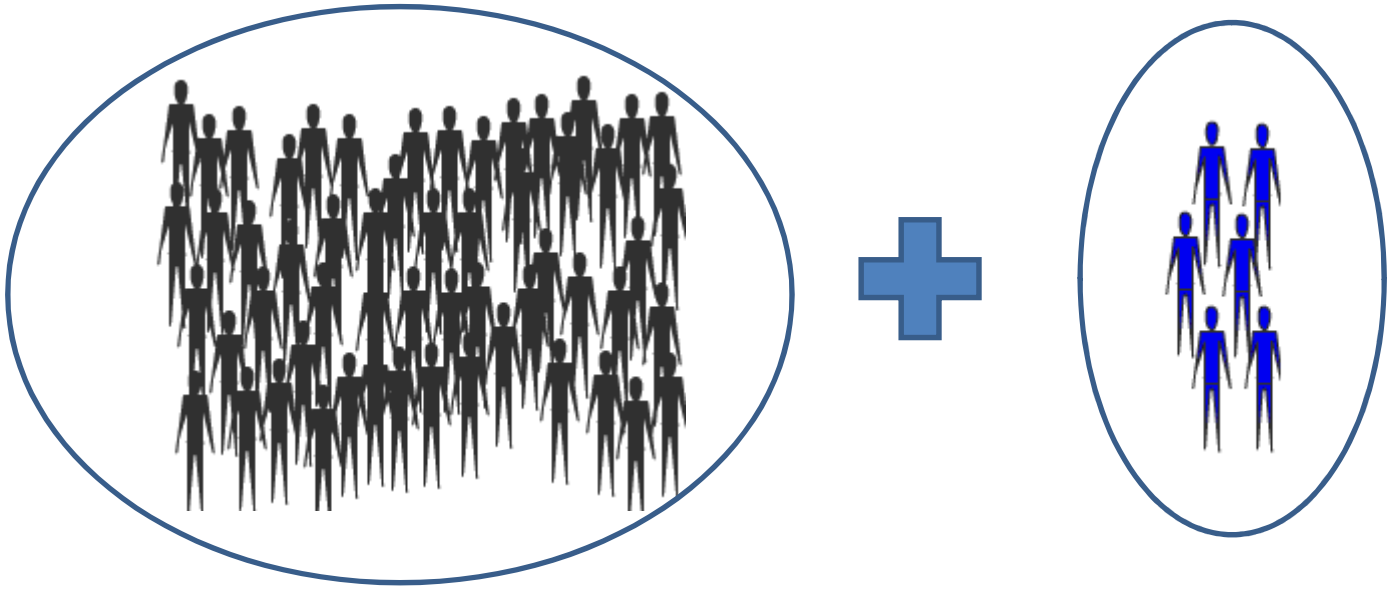
- Usually not possible to sequence the entire cohort
- Instead, most studies use small selected samples
- Sample ascertainment mechanism can be complicated, which may involve
  - Multiple phenotype
  - Extreme phenotypes
  - Family histories

Sequencing Based Genetic Studies

- Most studies are not well powered to detect associations for complex primary phenotypes
  - Kryukov et al 2009 *PNAS*

- In addition to the primary phenotype, many clinically important secondary phenotypes are often measured
  - Example:
    - BMI,
    - Diastolic and systolic blood pressure,
    - Blood cholesterol levels

Combine Samples from Different Studies for Mapping Secondary Phenotypes

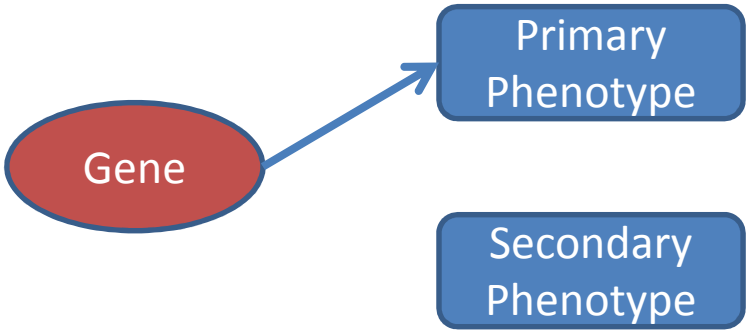


Phenome Mapping

- Combining multiple cohorts
- Different Cohorts may be collected for different primary phenotypes
- Joint analysis of shared primary or secondary phenotypes between different cohorts

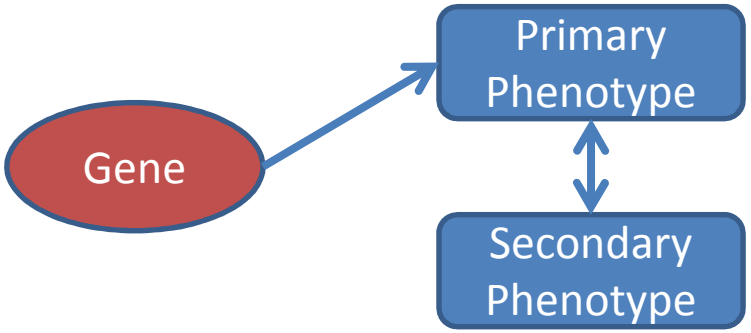
Mapping Secondary Phenotypes in Selected Samples

- The analysis of secondary phenotype can be biased in selected samples
  - if the sampling mechanism is not properly modeled
- Example: case-control study



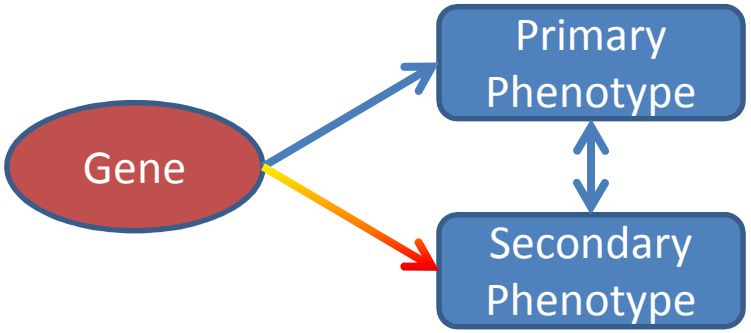
Mapping Secondary Phenotypes in Selected Samples

- The analysis of secondary phenotype can be biased in selected samples
  - if the sampling mechanism is not properly modeled
- Example: case-control study



Mapping Secondary Phenotypes in Selected Samples

- The analysis of secondary phenotype can be biased in selected samples
  - if the sampling mechanism is not properly modeled
- Example: case-control study



Mapping Secondary Phenotypes in Selected Samples I

- Methods developed for correcting the bias for case-control studies
  - Bias was evaluated
    - Kraft et al *Genetic Epidemiology*
  - Inverse sampling probability weighted regression
    - Richardson et al *Epidemiology 2007*
  - Maximum likelihood based approach
    - Lin and Zeng *Genetic Epidemiology 2009*

Maximum Likelihood Method

- Joint modeling of two phenotypes

$$\begin{cases} \log\left(\dfrac{P\left(Y_{i1}^{*}=1\right)}{1-P\left(Y_{i1}^{*}=1\right)}\right)=\beta_{01}+\beta_{1}X_{i}+\sum_{k}\alpha_{1k}W_{ik}+\gamma_{2}Y_{i2} \\ Y_{i2}=\beta_{02}+\beta_{2}X_{i}+\sum_{j}\alpha_{2k}W_{ik}+\varepsilon_{i2} \end{cases}$$

- Retrospective likelihood

$$L\left(\vec{\beta};\{X_{i},Y_{i1},Y_{i2}\}_{i}\right)=\prod_{i}P\left(Y_{i2}^{*},X_{i}\mid Y_{i1}\right)$$

Mapping Secondary Phenotypes in Selected Sample II

- Limitations of existing methods:
  - Developed for case control studies
  - Not directly applicable to more complicated ascertainment mechanisms
    - Especially when secondary phenotypes are also involved in sample ascertainment

Pleio-MAP method

- Multivariate liability threshold model

$Y_{i1}, Y_{i2} \sim$ liability traits for the primary and secondary phenotypes

$$\begin{cases} Y_{i1} = \beta_{01} + \beta_1 X_i + \sum_k \alpha_{1k} W_{ik} + \varepsilon_{i1} \\ Y_{i2} = \beta_{02} + \beta_2 X_i + \sum_j \alpha_{2k} W_{ik} + \varepsilon_{i2} \end{cases}$$

Pleio-MAP method

- Multivariate liability threshold model

$X_i \sim \text{locus genotype coding}$

$$\begin{cases} Y_{i1} = \beta_{01} + \beta_1 \boxed{X_i} + \sum_k \alpha_{1k} W_{ik} + \varepsilon_{i1} \\ Y_{i2} = \beta_{02} + \beta_2 \boxed{X_i} + \sum_j \alpha_{2k} W_{ik} + \varepsilon_{i2} \end{cases}$$

Pleio-MAP method

- Multivariate liability threshold model

W_{ik} ~ covariates for individual i

$$\begin{cases} Y_{i1} = \beta_{01} + \beta_1 X_i + \sum_k \alpha_{1k} \boxed{W_{ik}} + \varepsilon_{i1} \\ Y_{i2} = \beta_{02} + \beta_2 X_i + \sum_j \alpha_{2k} \boxed{W_{ik}} + \varepsilon_{i2} \end{cases}$$

Pleio-MAP method

- Multivariate liability threshold model

$$(\varepsilon_{i1}, \varepsilon_{i2}) \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

$$\begin{cases} Y_{i1} = \beta_{01} + \beta_1 X_i + \sum_k \alpha_{1k} W_{ik} + \boxed{\varepsilon_{i1}} \\ Y_{i2} = \beta_{02} + \beta_2 X_i + \sum_j \alpha_{2k} W_{ik} + \boxed{\varepsilon_{i2}} \end{cases}$$

Pleio-MAP method

- For a multivariate liability threshold model (MLT)
  - Liability trait may not be directly observed
    - If observed, MLT is equivalent to a multivariate normal model
  - A binary (or ordinal) phenotype may be observed, e.g.

$$Y_{i1}^{*} = \begin{cases} 1 & Y_{i1} > y_1^C \\ 0 & Y_{i1} \leq y_1^C \end{cases}$$

Pleio-MAP method

- Applications to selected samples:
  - Jointly model sampling mechanism and correlations between multiple phenotypes
  - Prospective likelihood approach
    - Joint probability of multiple phenotypes conditional on the sampling scheme

Ascertainment Status

$$L(\beta,\theta;X,\mathrm{Y})=\prod_{i=1}^{N_U+N_A} p\left(Y_{i1},Y_{i2},X_i \middle| \boxed{Z_i=1},\{W_{ik}\}_k\right)$$

Pleio-MAP method

- Modeling of sampling schemes

Sampling mechanism

$$p\left(Y_{i1}, Y_{i2}, X_i \middle| Z_i = 1, \{W_{ik}\}_k\right) = \frac{\boxed{P\left(Z_i = 1 \middle| Y_{i1}, Y_{i2}, X_i, \{W_{ik}\}_k\right)} p\left(Y_{i1}, Y_{i2}, X_i \middle| \{W_{ik}\}_k\right)}{\int P\left(Z_i = 1 \middle| y_{i1}, y_{i2}\right) p\left(y_{i1}, y_{i2}\right) dy_{i1} dy_{i2}}$$

- Pleio-MAP is applicable to study designs for which the sampling scheme can be modeled

PLeio-MAP method

- For a case control study

$$P\left(Z_i = 1 \middle| Y_{i1}, Y_{i2}, X_i, \{W_{ik}\}_k\right) = P\left(Z_i = 1 \middle| Y_{i1}\right)$$

- The sampling probability should satisfiy

$$\frac{P\left(Z_i = 1 \middle| Y_{i1} = 1\right)}{P\left(Z_i = 1 \middle| Y_{i1} = 0\right)} = \frac{N^A P\left(Y_{i1} = 0\right)}{N^U P\left(Y_{i1} = 1\right)}$$

Pleio-MAP method

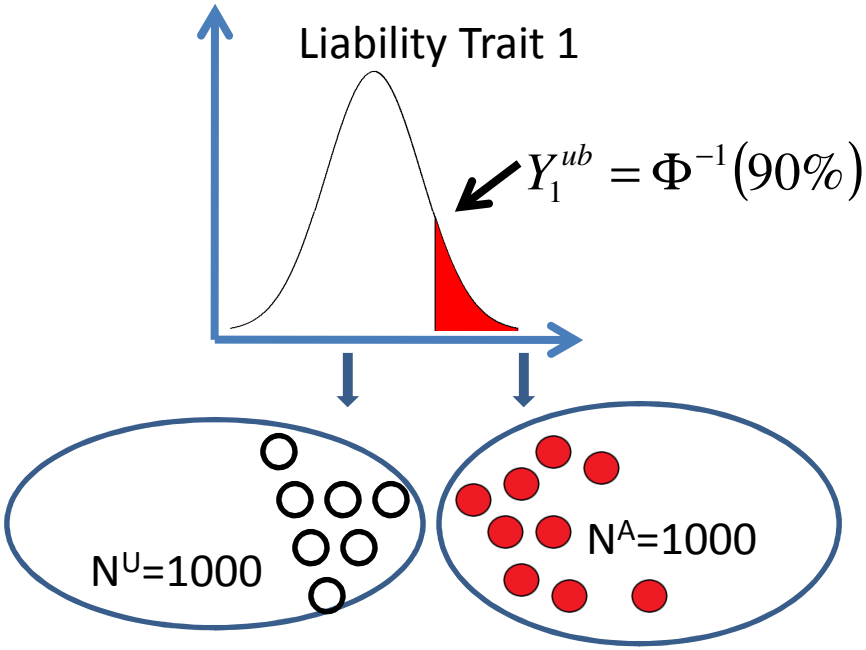
- Testing of Associations:
  - Likelihood based tests:
    - Likelihood ratio test
    - Score test
    - Wald test
- Combine multiple cohort
  - Test of heterogeneity
  - Combine individual participants data
  - Combine estimates of genetic effects using meta-analysis based approach

Simulation Experiment 1

Comparisons of Different Selective Sampling Designs for Mapping Secondary Phenotypes

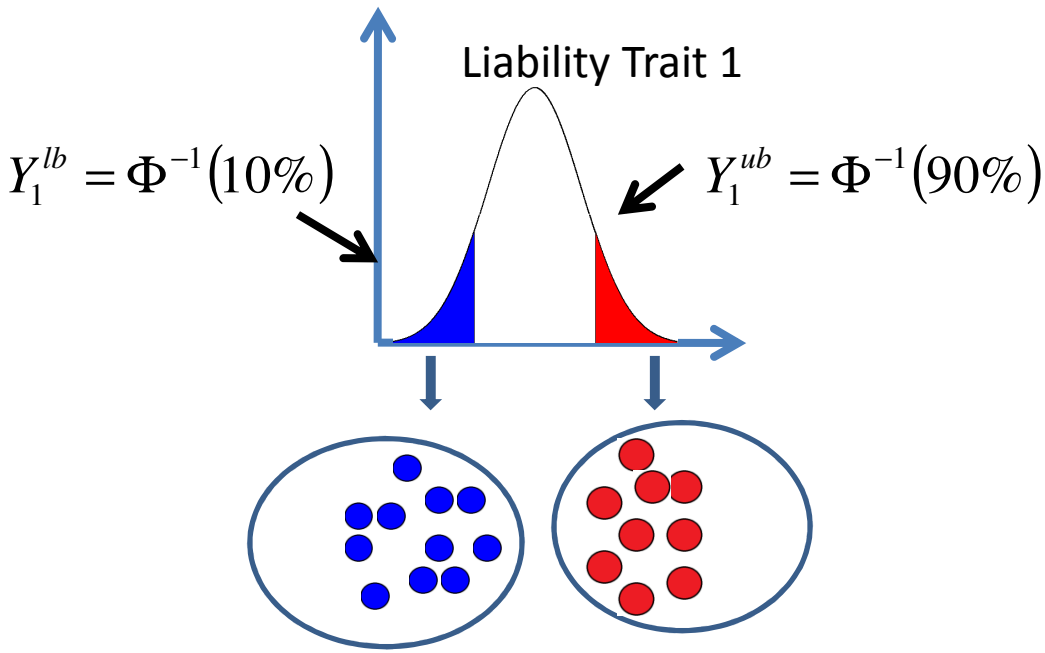
Study Designs in Sequencing Based Genetic Studies

Case Control Design



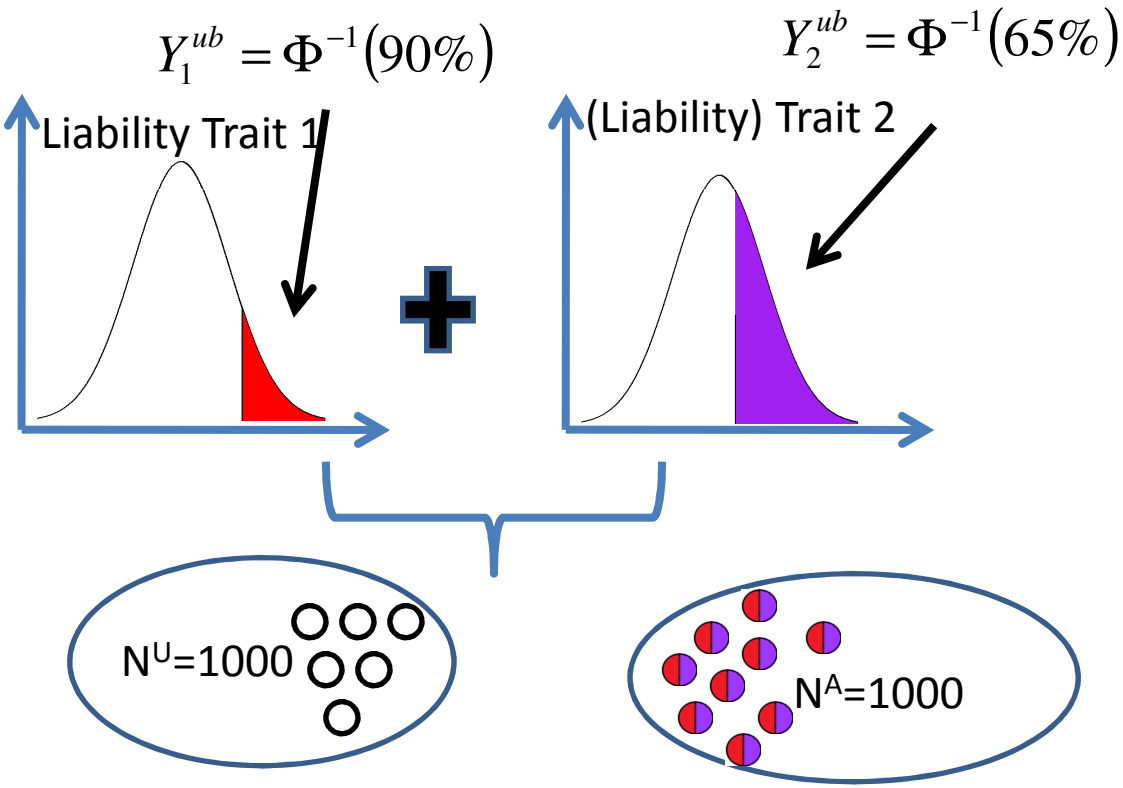
Study Designs in Sequencing Based Genetic Studies

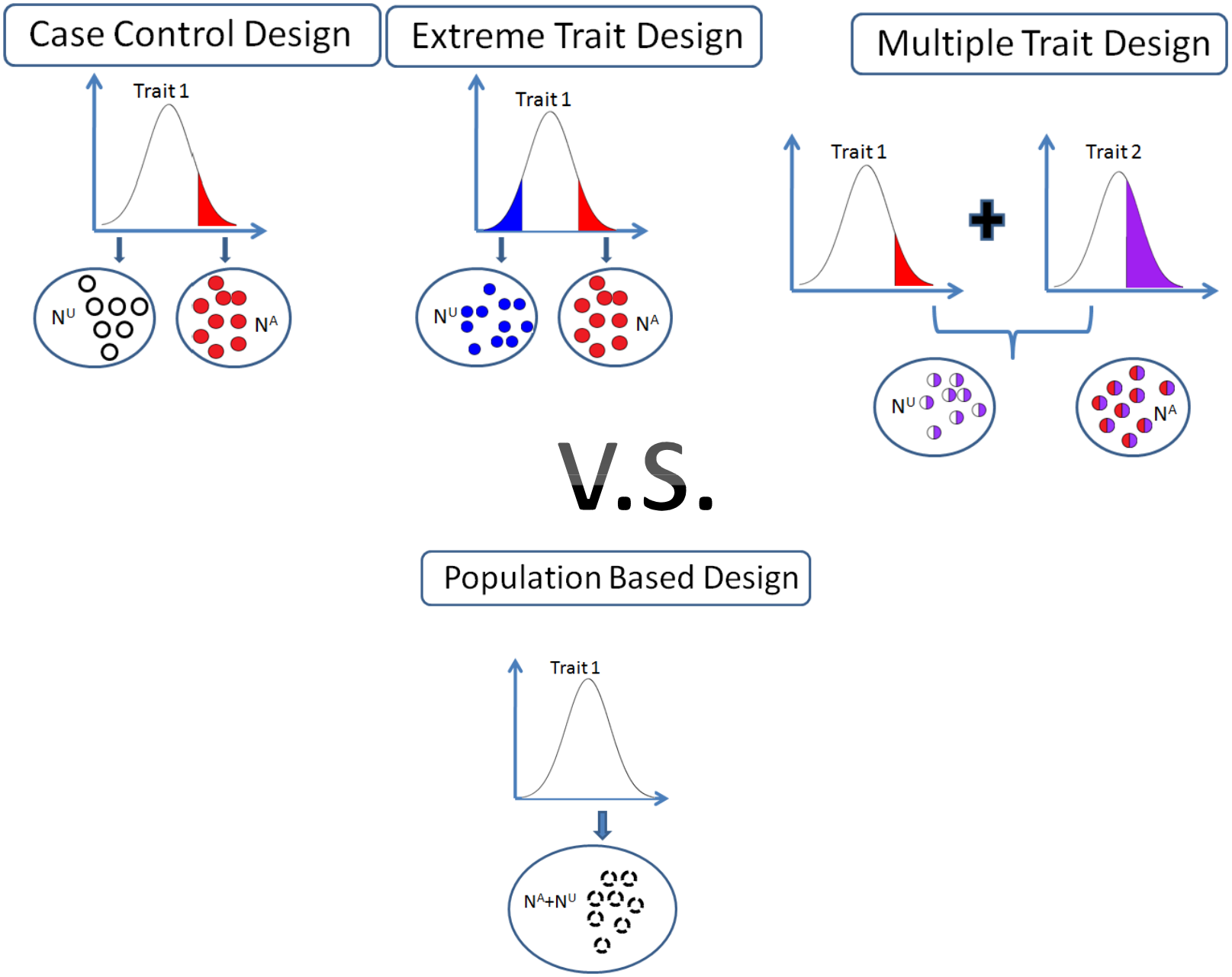
Extreme Trait Design



Study Designs in Sequencing Based Genetic Studies

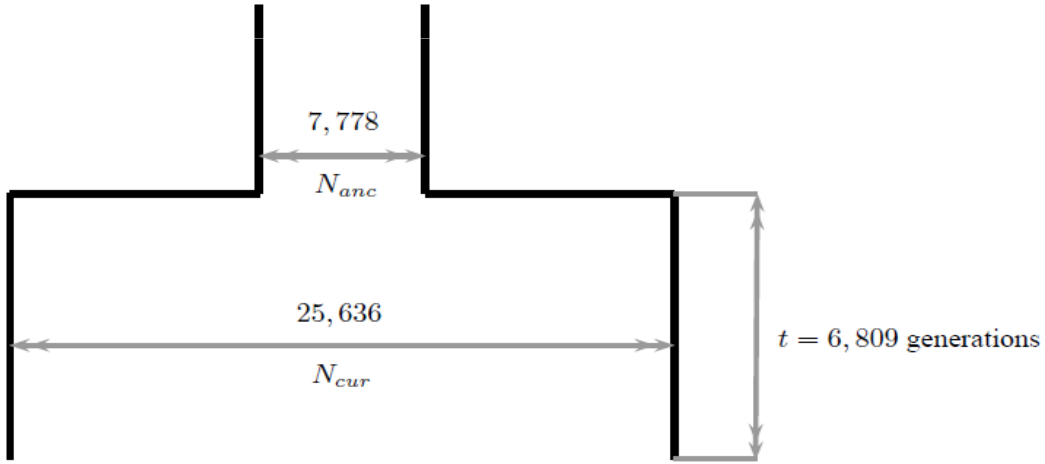
Multiple Trait Design




Case Control Design
Trait 1
N^U
N^A
Extreme Trait Design
Trait 1
N^U
N^A
Multiple Trait Design
Trait 1
Trait 2
N^U
N^A
V.S.
Population Based Design
Trait 1
N^A+N^U

Simulations of Genetic Data

- Demographic change of Africans:
  - Boyko et al *PLoS Genetics 2008*



Simulations of Genetic Data

- Purifying selections:
  - Selective disadvantage of new mutations
    - Heterozygous u
    - Homozygous $2u$
  - Scaled disadvantage: $\gamma = 2N_{curr}u$

$$\gamma = -x, x \sim \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx),$$

  - where

$$a = 0.184, b = 8{,}200$$

Simulations of Phenotype Data

- Multi-site non-synonymous variants genotype for individual "*i*"

$$\vec{X}_i = \left(x_i^1, \cdots, x_i^S\right)$$

- Among the S non-synonymous variant sites,
  - A subset of variant sites C_1 are randomly selected as causative variant sites for liability trait 1
  - Another subset of variant sites C_2 are independently selected as causative variant sites for trait 2

Simulations of Phenotype Data

- Two liability traits are generated according to

$$\left(Y_{i1}, Y_{i2}\right) \sim \mathrm{N}\left(\vec{\mu}_i, \Sigma\right)$$

where $\vec{\mu}_i = \left(\tilde{\beta}_1 \sum_{s \in C_1} x_i^s, \beta_0 + \tilde{\beta}_2 \sum_{s \in C_2} x_i^s\right), \Sigma = \begin{pmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{pmatrix}$

Simulation of Phenotype Data

- Choice of parameters
  - For the liability trait 1:
    - $\beta_1 = 0 \text{ or } \beta_1 = 0.5$
  - For liability trait 2:
    - $\beta_2 = -0.5\sigma_2 \text{ or } \beta_2 = 0.5\sigma_2$
  - Phenotypic residual correlations:
    - $\rho = 0.6 \text{ or } \rho = -0.6$
- 1,000 individuals sequenced
- Significance level $\alpha = 0.05$

Power for Case Control Study

Genetic Parameters			Power[d]
$\tilde{\beta}_1$	$\tilde{\beta}_2$	ρ	
0.5	-0.5	-0.6	0.533
0.5	-0.5	0.6	0.556
0.5	0.5	-0.6	0.565
0.5	0.5	0.6	0.545
0	-0.5	-0.6	0.527
0	-0.5	0.6	0.513
0	0.5	-0.6	0.521
0	0.5	0.6	0.531

Power for Population Based Study

Sample Size	1000	2000	3000
Power	0.516	0.666	0.736

Power for Extreme Trait Study

Genetic Parameters			Power[d]
$\tilde{\beta}_1$	$\tilde{\beta}_2$	ρ	
0.5	-0.5	-0.6	0.582
0.5	-0.5	0.6	0.654
0.5	0.5	-0.6	0.667
0.5	0.5	0.6	0.589
0	-0.5	-0.6	0.598
0	-0.5	0.6	0.609
0	0.5	-0.6	0.606
0	0.5	0.6	0.602

Power for Population Based Study

Sample Size	1000	2000	3000
Power	0.516	0.666	0.736

Power for Multi-trait Study

Genetic Parameters			Power[d]
$\tilde{\beta}_1$	$\tilde{\beta}_2$	ρ	
0.5	-0.5	-0.6	0.292
0.5	-0.5	0.6	0.471
0.5	0.5	-0.6	0.391
0.5	0.5	0.6	0.562
0	-0.5	-0.6	0.315
0	-0.5	0.6	0.447
0	0.5	-0.6	0.373
0	0.5	0.6	0.549

Power for Population Based Study

Sample Size	1000	2000	3000
Power	0.516	0.666	0.736

Results for Experiment 1

- Analyzing secondary phenotypes in selected samples can be more powerful than population based unselected samples
  - Although it was believed that population sample is suitable for mapping multiple phenotypes

Results for Experiment 1

- This is because:
  - Variants with pleiotropic effects will be enriched in the selected sample
  - Due to phenotypic correlations, selections through primary phenotype induce selections on the secondary phenotype

Simulation Experiment 2

Combining Case Control Study and Multiple Trait Study

Results for Combining Multiple Studies

Parameters						Power[g]		
$\tilde{\beta}_A^{CC}$	$\tilde{\beta}_T^{CC}$	$\rho_{C,T}^{CC}$	$\tilde{\beta}_C^{MT}$	$\tilde{\beta}_T^{MT}$	$\rho_{C,T}^{MT}$	Case Control Design	Mutlple - phenotype Design	Meta-Analysis
0	-0.5	-0.3	0.5	-0.5	0.3	0.510	0.418	0.690
0	-0.5	0.3	0.5	-0.5	0.3	0.499	0.418	0.680
0	0.5	-0.3	0.5	0.5	0.3	0.508	0.526	0.726
0	0.5	0.3	0.5	0.5	0.3	0.517	0.526	0.732
0	-0.5	-0.6	0.5	-0.5	0.3	0.527	0.418	0.703
0	-0.5	0.6	0.5	-0.5	0.3	0.513	0.418	0.685
0	0.5	-0.6	0.5	0.5	0.3	0.521	0.526	0.731
0	0.5	0.6	0.5	0.5	0.3	0.531	0.526	0.741

Power for Population Based Study

Sample Size	1000	2000	3000
Power	0.516	0.666	0.736

Results for Combining Multiple Studies

Parameters						Power[g]		
$\tilde{\beta}_A^{CC}$	$\tilde{\beta}_T^{CC}$	$\rho_{C,T}^{CC}$	$\tilde{\beta}_C^{MT}$	$\tilde{\beta}_T^{MT}$	$\rho_{C,T}^{MT}$	Case Control Design	Mutlple - phenotype Design	Meta-Analysis
0	-0.125	-0.3	0.5	-0.5	0.3	0.091	0.418	0.444
0	-0.125	0.3	0.5	-0.5	0.3	0.106	0.418	0.459
0	0.125	-0.3	0.5	0.5	0.3	0.128	0.526	0.550
0	0.125	0.3	0.5	0.5	0.3	0.105	0.526	0.529
0	-0.125	-0.6	0.5	-0.5	0.3	0.097	0.418	0.426
0	-0.125	0.6	0.5	-0.5	0.3	0.117	0.418	0.462
0	0.125	-0.6	0.5	0.5	0.3	0.119	0.526	0.569
0	0.125	0.6	0.5	0.5	0.3	0.102	0.526	0.534

Power for Population Based Study

Sample Size	1000	2000	3000
Power	0.516	0.666	0.736

Analysis of ANGPTL 3,4,5 and 6 Genes

- Data generated by Dallas Heart Study (DHS)

- ANGPTL3,4,5 and 6 genes sequenced for a multi-ethnic population-based sample of 1830 African Americans, 1045 European Americans, 601 Hispanic Americans, 75 from other ethinicities

Analysis of ANGPTL 3,4,5 and 6 Genes

- Eight metabolism phenotypes are measured:
  - Body mass index (BMI)
  - Diastolic blood pressure (DiasBP)
  - Systolic blood pressure (SysBP)
  - Total cholesterol level (TCL)
  - Low density lipoprotein (LDL)
  - High density lipoprotein (HDL)
  - Triglyceride (TG)
  - Glucose (Gluc)

Results for Primary Trait Analysis

- Each phenotype was analyzed as the primary phenotype using individuals from the top and bottom quartile

Phenotypes		Estimates	
	p-values[a]	Locus Genetic Effect Estimates(σ_r)	Carrier Frequency[b]
ANGPTL3			
BMI	0.924	--	0.07
DiasBP	0.898	--	0.073
SysBP	0.997	--	0.069
TCL	0.253	--	0.063
LDL	0.974	--	0.067
HDL	0.733	--	0.068
TG	0.077	--	0.062
Gluc	0.640	--	0.071

Phenotypes		Estimates	
	p-values[a]	Locus Genetic Effect Estimates(σ_r)	Carrier Frequency[b]
ANGPTL4			
BMI	0.504	--	0.096
DiasBP	0.608	--	0.082
SysBP	0.679	--	0.094
TCL	0.311	--	0.09
LDL	0.179	--	0.086
HDL	0.068	--	0.093
TG	0.005*	-0.195	0.086
Gluc	0.541	--	0.101

Phenotypes		Estimates	
	p-values[a]	Locus Genetic Effect Estimates(σ_r)	Carrier Frequency[b]
ANGPTL5			
BMI	0.003*	0.215	0.095
DiasBP	0.564	--	0.1
SysBP	0.842	--	0.108
TCL	0.355	--	0.096
LDL	0.600	--	0.102
HDL	0.024*	0.151	0.097
TG	0.894	--	0.095
Gluc	0.665	--	0.105

Phenotypes		Estimates	
	p-values[a]	Locus Genetic Effect Estimates(σ_r)	Carrier Frequency[b]
ANGPTL6			
BMI	0.022*	0.219	0.051
DiasBP	0.110	--	0.057
SysBP	0.487	--	0.051
TCL	0.479	--	0.051
LDL	0.628	--	0.055
HDL	0.431	--	0.053
TG	0.978	--	0.049
Gluc	0.205	--	0.05

Results

- Each additional phenotype was analyzed as secondary phenotype

Primary Phenotype	P-values for Analysis of Secondary Phenotypes[a]							
	BMI	DiasBP	SysBP	TCL	LDL	HDL	TG	Gluc
	ANGPTL 3							
BMI		0.649	0.766	0.429	0.681	0.717	0.121	0.114
DiasBP	0.941	-	0.889	0.580	0.745	0.309	0.441	0.398
SysBP	0.550	0.509	-	0.371	0.223	0.689	0.073	0.222
TCL	0.988	0.955	0.327	-	0.971	0.289	0.163	0.151
LDL	0.871	0.372	0.349	0.114	-	0.116	0.183	0.024*
HDL	0.945	0.616	0.312	0.825	0.668	-	0.561	0.639
TG	0.910	0.883	0.437	0.945	0.418	0.863	-	0.148
Gluc	0.652	0.208	0.351	0.982	0.475	0.692	0.335	-
	ANGPTL 4							
BMI	-	0.292	0.268	0.733	0.440	0.497	0.025*	0.972
DiasBP	0.965	-	0.380	0.361	0.363	0.121	0.137	0.389
SysBP	0.993	0.551	-	0.728	0.754	0.099	0.012*	0.405
TCL	0.861	0.532	0.571	-	0.052	0.759	0.065	0.933
LDL	0.281	0.894	0.269	0.135	-	0.053	0.010*	0.999
HDL	0.708	0.904	0.286	0.318	0.262	-	0.107	0.874
TG	0.310	0.364	0.584	0.629	0.326	0.784	-	0.845
Gluc	0.824	0.524	0.084	0.848	0.561	0.479	0.118	-
	ANGPTL 5							
BMI	-	0.920	0.114	0.521	0.233	0.056	0.377	0.797
DiasBP	0.118	-	0.096	0.451	0.803	0.092	0.616	0.367
SysBP	0.203	0.887	-	0.117	0.160	0.304	0.791	0.294
TCL	0.107	0.536	0.923	-	0.399	0.014*	0.221	0.488
LDL	0.084	0.735	0.587	0.202	-	0.002*	0.147	0.458
HDL	0.387	0.866	0.917	0.463	0.991	-	0.569	0.900
TG	0.044*	0.871	0.074	0.296	0.597	0.185	-	0.448
Gluc	0.030*	0.779	0.957	0.546	0.717	0.002*	0.451	-
	ANGPTL 6							
BMI	-	0.300	1.000	0.606	0.457	0.324	0.401	0.419
DiasBP	0.008*	-	0.385	0.459	0.690	0.478	0.721	0.197
SysBP	0.773	0.816	-	0.622	0.853	0.668	0.338	0.490
TCL	0.024*	0.530	0.992	-	0.823	0.324	0.702	0.940
LDL	0.089	0.383	0.850	0.485	-	0.429	0.801	0.314
HDL	0.034*	0.101	0.873	0.800	0.870	-	0.393	0.215
TG	0.210	0.735	0.974	0.357	0.695	0.561	-	0.811
Gluc	0.153	0.402	0.897	0.340	0.531	0.267	0.905	-

Conclusions

- Mapping secondary phenotypes in selected samples is possible

- There is considerable power to detect secondary phenotype associations in selected samples

Conclusions

- Report estimates of genetic effects for multiple traits

- Collect and use well phenotyped cohort
  - Measure relevant phenotypes in addition to the primary phenotype
    - Missing phenotypes are hard to “impute”
  - Record sample ascertainment mechanisms

Acknowledgement

- Prof. Suzanne Leal
- Gao Wang
- Drs. Jonathan Cohen and Helen Hobbs for sharing sequence data from *ANGPTL3*, *ANGPTL4*, *ANGPTL5* and *ANGPTL6* genes.
- Computation for this research was supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems and Sigma Solutions, Inc