Integrating Multiple Correlated Phenotypes for Genetic Association Analysis by Maximizing Heritability

Jin Zhou

Michael Cho, Christoph Lange, Sharon Lutz,
Edwin Silverman, and Nan Laird

Division of Epidemiology and Biostatistics
University of Arizona

ICSA 2013

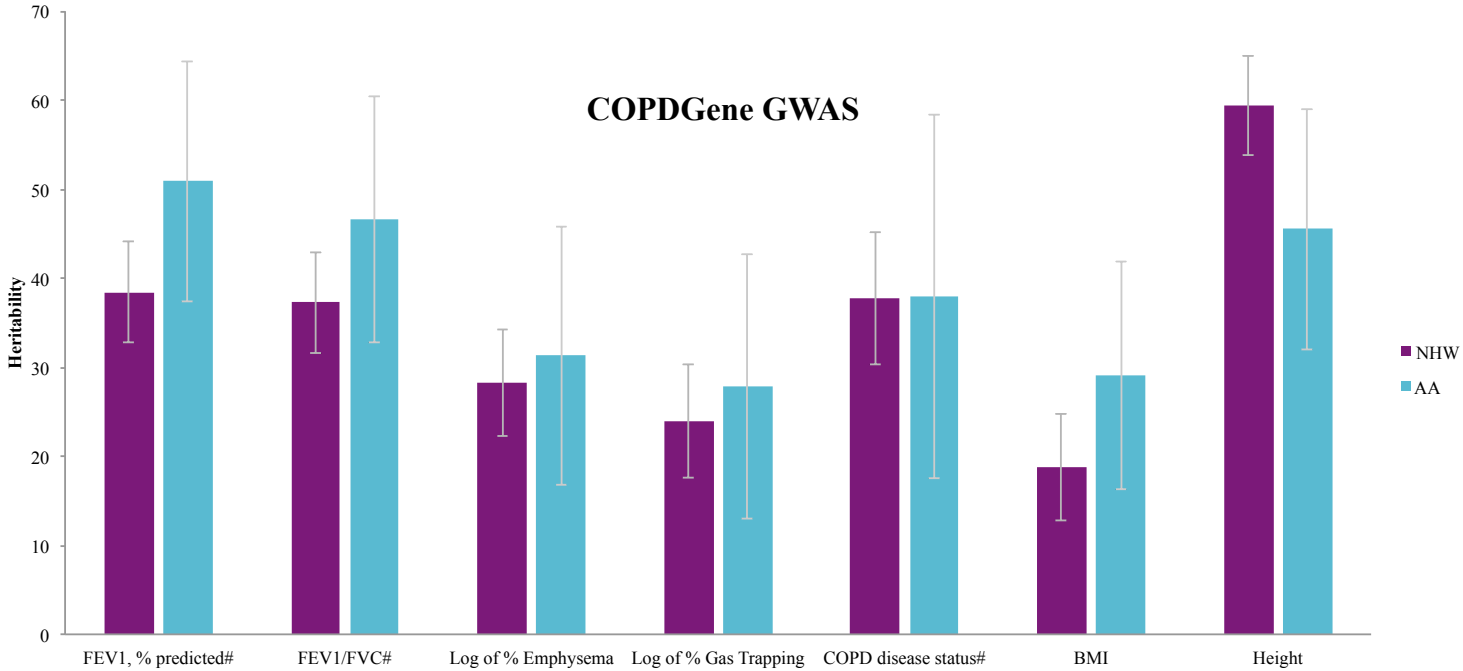
Motivation

Chronic Obstructive Pulmonary Disease (COPD) as an example.

http://www.copdgene.org/

- Cross sectional prospective cohort study (2008-2011).
- Non-Hispanic While (NWH): 7000 subjects.
- African American (AA): 3000 subjects.
- Lung function phenotypes: FEV_1, FEV_1/FVC; imaging phenotypes: emphysema, gastrapping.

Heritability Estimation



[Zhou et al., AJRCCM, 2013]

Pleiotropy Effects of COPD-Related Phenotypes

Pleiotropy: one gene influences multiple phenotypic traits.

Genotypic correlation				
	FEV1	FEV1/FVC	Log(% Emph)	Log(% GasTrap)
FEV1	–	0.889	-0.626	-0.844
FEV1/FVC	0.797	–	-0.818	-0.877
Log(% Emph)	-0.573	-0.725	–	0.903
Log(% GasTrap)	-0.855	-0.696	0.814	–

[Zhou et al., AJRCCM, 2013]

Multiple Phenotype Association Analysis

- Multivariate analysis (LMM or GEE).
- Combining test statistics.
- Principle component analysis.

Multiple Phenotype Integration

- Y denote the p dimensional vector of phenotypes, $Y = (Y_1, ..., Y_p)$, where each Y_i is an n by 1 vector.
- The heritability of a linear combination of phenotypes $l'Y$, can be expressed as

$$h_l^2 = \frac{lV_g l}{lV_p l},$$

  where V_g and V_p are genetic and phenotypic variance-covariance matrix.
- The goal is to find a set of coefficients $l = (l_1, \ldots, l_p)$ such that $l'Y$ has the maximum heritability.

MaxH Phenotype

Optimization problem can by solved by eigen-decomposition,

- The first linear combination coefficients l_1 is the eigenvector corresponding to the biggest eigenvalue of the generalized eigen-system,

$$V_g l_1 \quad = \quad \lambda V_p l_1 .$$

Heritability of combined phenotype is determined by,

- The structure of the genotypic and phenotypic variance-covariance matrix;
- The number of phenotypes combined.

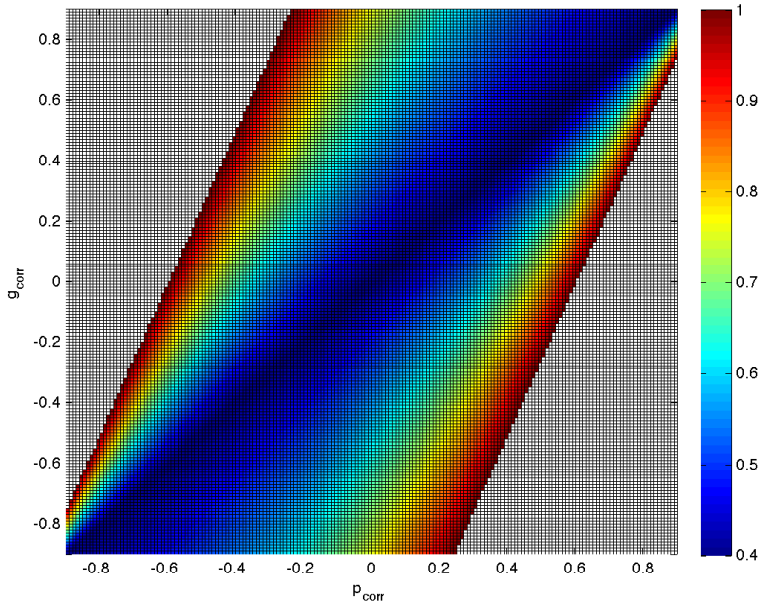
Combing Two Phenotypes

Genotypic and phenotypic variance-covariance matrices take form,

$$V_g = h^2 \begin{pmatrix} 1 & r_g k \\ r_g k & k^2 \end{pmatrix} \quad V_p = \begin{pmatrix} 1 & r_p \\ r_p & 1 \end{pmatrix}$$

where r_g and r_p are the genotypic and phenotypic correlation coefficients. If $k = 1$,

- $r_g > r_p$, $h^2_{\mathrm{m}} = \frac{1+r_g}{1+r_p} h^2$ and $l = (1, 1)$.
- $r_g < r_p$, $h^2_{\mathrm{m}} = \frac{1-r_g}{1-r_p} h^2$ and $l = (1, -1)$.
- $r_g = r_p$, $h^2_{\mathrm{m}} = h^2$



Compare with PCA

When $r_p > 0$, first eigenvector of PCA is $(1,1)$.

- If $r_g > r_p$ PCA is the same as MaxH approach.
- If $r_g < r_p$ PCA still takes vector $(1,1)$ while the other eigenvector $(1,-1)$ leads to the maximized heritability.

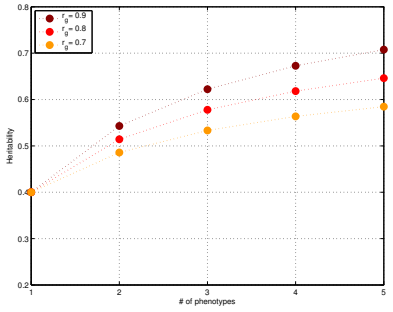
This result can be generalized to combing T standardized phenotypes, each pair with common r_g and r_p.

Combing More Than Two Phenotypes

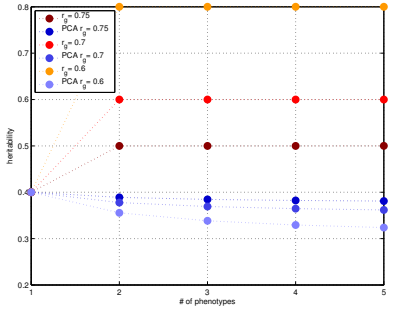
$$V_g = h^2 \begin{pmatrix} 1 & kr_g & \dots & k^t r_g \\ kr_g & k^2 & \dots & k^{t+1} r_g \\ \dots & \dots & \dots & \dots \\ k^{t+1} r_g & \dots & k^{2(t-1)} & k^{2t-1} r_g \\ k^t r_g & \dots & k^{2t-1} r_g & k^{2t} \end{pmatrix}$$

$$V_p = \begin{pmatrix} 1 & r_p & \dots & r_p \\ r_p & 1 & \dots & r_p \\ \dots & \dots & \dots & \dots \\ r_p & \dots & 1 & r_p \\ r_p & \dots & r_p & 1 \end{pmatrix}.$$

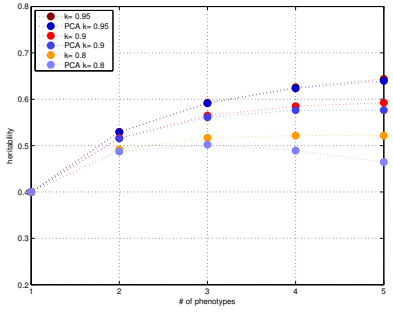
$(a) : r_p = 0.4, k = 1$



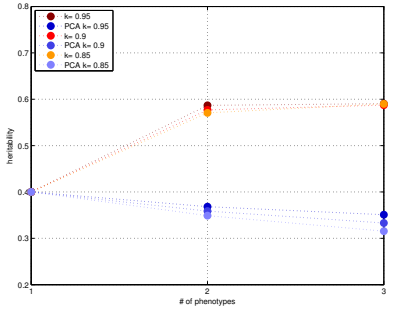
$(b) : r_p = 0.8, k = 1$



$(c) : r_p = 0.4, r_g = 0.9$



$(d) : r_p = 0.8, r_g = 0.7$



Empirical Power and Parameter Estimates

- Two phenotypes were simulated based on the linear model assuming 40% of the heritability.
- 100 SNPs among Chromosome 1 (taken from COPDGene NWH) were randomly chosen as the causal SNPs for polygenic background.
- One SNP with MAF 0.22 was chosen to explain 2% of the total heritability.
- 500 replicates, each with 3000 individuals. Significant level 5×10^{-5}.

	Power				Het. Est.			
	$r_g = 0.9$, $r_p = 0.4$		$r_g = 0.7$, $r_p = 0.8$		$r_g = 0.9$, $r_p = 0.4$		$r_g = 0.7$, $r_p = 0.8$	
True	MaxH	0.492	MaxH	0.524	MaxH	0.506(0.049)	MaxH	0.566(0.113)
	PCA	0.492	PCA	0.396	PCA	0.506(0.049)	PCA	0.352(0.048)
Est.	MaxH	0.493	MaxH	0.492	MaxH	0.509(0.049)	MaxH	0.573(0.110)
	PCA	0.492	PCA	0.402	PCA	0.506(0.049)	PCA	0.353(0.051)
Trait 1	0.434		0.398		0.365(0.043)		0.372(0.048)	
Trait 2	0.412		0.378		0.374(0.049)		0.371(0.049)	

GWAS analysis in COPDGene NWH population

- Phenotypes: FEV_1, FEV_1/FVC and log(percent Emphysema).
- Covariates: Sex, AgeEnroll, BMI, ATS_PackYears, PC1-5, SmokCigNow.
- SNPs passing genome-wide significant level 5×10^{-7}.

Chr	Nearest Gene	MaxH	PCA	Multivariate
1	*TGFB2*			2
4	*FAM13A*	2		
4	*HHIP*	6	7	5
6	*AGER*		1	1
11	*MMP12*	1	1	
15	*CHRNA3/CHRNA5/AGPHD1/IREB2*	13	15	13
18	*PTPRM*			1

Acknowledgments

- Harvard School of Public Health
  - Nan Laird
  - Christoph Lange
- Channing's Lab at Harvard Medical School
  - Edwin Silverman
  - Michael Cho
- University of Colorado
  - Sharon Lutz

Power of Association Analysis

Determined by Non-Centrality Parameter. Assume we test a SNP with effects that explain a fraction c $(0 \leq c \leq 1)$ of the total heritability, the NCP can be written as

$$\delta = N\frac{ch^2}{1-ch^2}.$$

Genetic Covariance

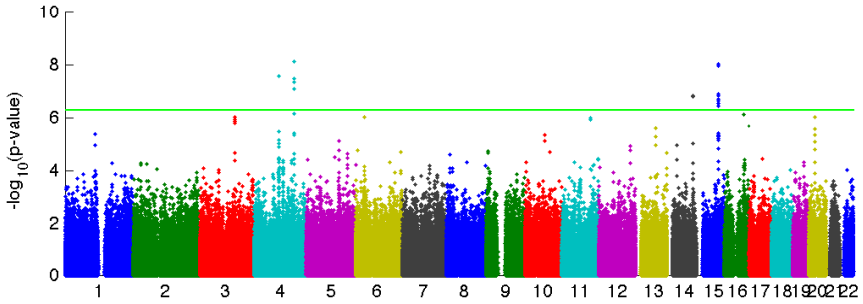
- Need to estimate V_g.
- Maximize likelihood.
- Method of moments. Suppose $Z = Y_1 + Y_2$,

$$\begin{aligned} \mathrm{Var}(Z) &= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\,\mathrm{Cov}(Y_1, Y_2) \\ \mathrm{Cov}(Y_1, Y_2) &= \frac{1}{2}[\mathrm{Var}(Z) - \mathrm{Var}(Y_1) - \mathrm{Var}(Y_2)] \end{aligned}$$

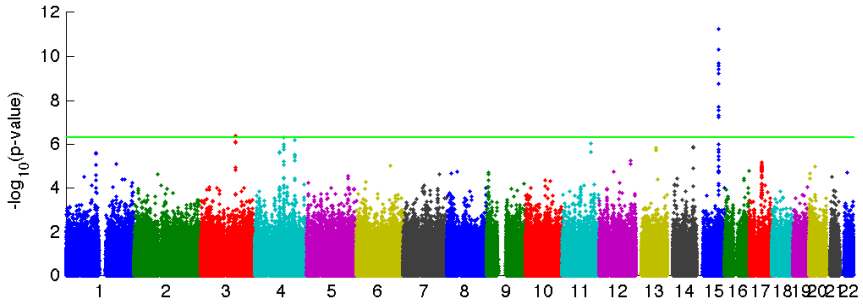
Pulmonary Phenotypes

MaxH=0.751Fev + 0.053FevFvc + 0.657pre FevFvc

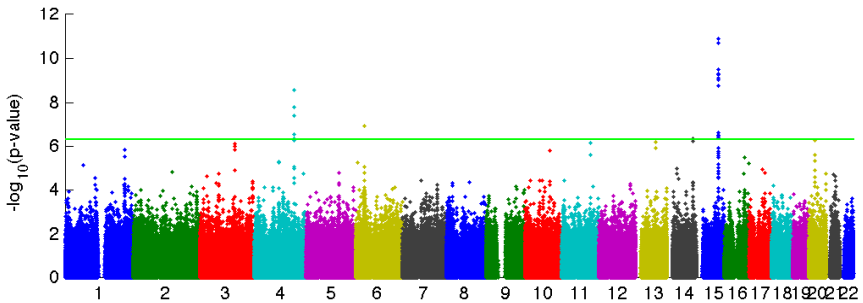
MaxH ($h^2 = 0.393$)



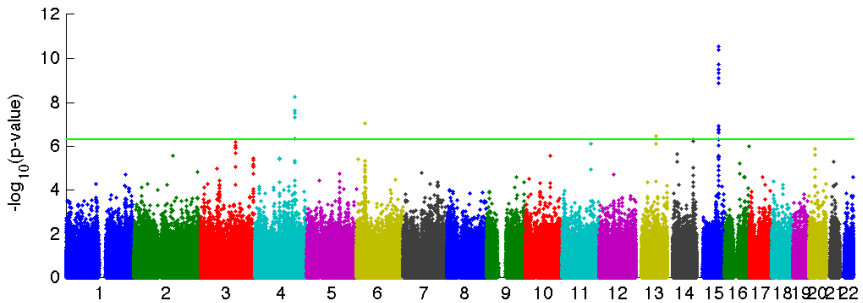
Fev ($h^2 = 0.383$)



FevFvc ($h^2 = 0.372$)



Pre FevFvc ($h^2 = 0.381$)



Impact of Missing Data

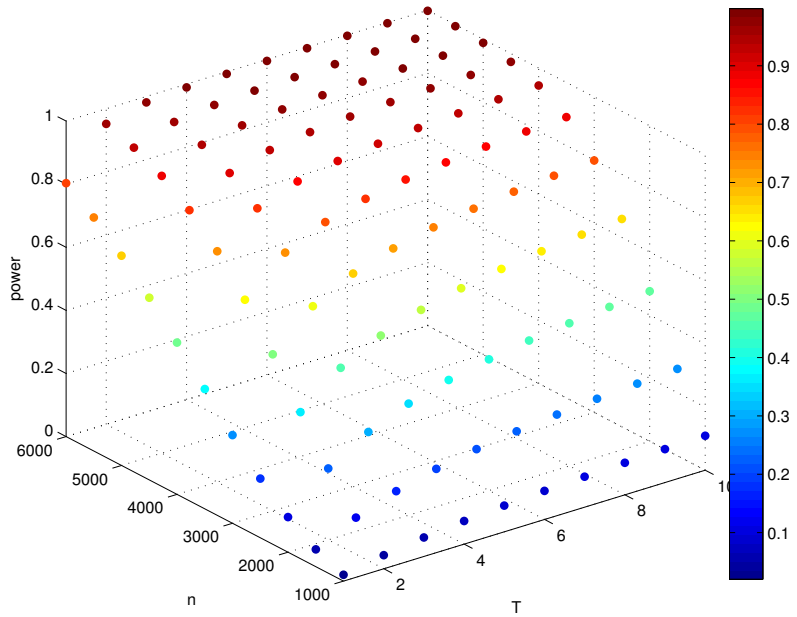


Figure : Power as a function of sample sizes where $r_p = 0.4$, $r_g = 0.9$, $k = 1$, $h^2 = 0.4$.