Genomic Privacy Challenges when Sharing Quantitative Trait GWAS Results

Hae Kyung Im

THE UNIVERSITY OF CHICAGO

August 22, 2014

Genomic Data Surge Since First Draft of Human Genome

- Biomedical research revolutionized by massive amounts of genomic data
- Huge potential for new discoveries
- "Few blockbuster new cures" (NYTimes)
- For full advantage, broad sharing of data and results is needed
- However, privacy of study participants has to be protected

Challenges in Sharing Genomic Results

- Summary statistics in large studies considered safe to publish
  proportion of females vs. males,
  average LDL cholesterol levels, etc.
- Genome wide association studies GWAS
  - for millions of SNPs
    differential mutations frequencies in cases vs. controls are generated
- Frequency of mutations in cases and controls used to be publicly available

Forensic Study Revealed Vulnerability

- Forensic application - Homer et al (2008) Plos Genetics
- Efficiency of new genotyping chips in forensic application
  - DNA sample from crime scene
  - DNA from suspect
  - Determine whether suspect's DNA is part of the sample

	Id 1	Id 2	Id 3	Id 4	Sample	Popul	Suspect
SNP 1	1	2	0	0	0.75	1.10	0
SNP 2	1	0	0	1	0.50	1.25	1
	...	...	...	...	...	...	...
SNP M	1	0	1	2	1.00	1.50	2

- Implication for GWAS results
- NIH withdrew public access to aggregate results

Quantitative Trait GWAS - What Are the Risks of Sharing?

- Quantitative Trait GWAS
  - $Y_i = \alpha_j + \beta_j X_{i,j} + e_i$
  - $\hat{\beta}_j = \left(\tilde{\mathbf{X}}_j' \tilde{\mathbf{X}}_j\right)^{-1} \tilde{\mathbf{X}}_j' \tilde{\mathbf{Y}}$
- My colleagues wanted to publish regression coefficients for studies in dbGaP
  but wanted a mathematical proof that re-identification was not possible

Betas and Genotype Are Known

$\hat{\beta}_1$	$X_{I,1}$
$\hat{\beta}_2$	$X_{I,2}$
$\vdots$	$\vdots$
$\hat{\beta}_M$	$X_{I,M}$

Average the product

$$\frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_j X_{I,j}$$

Testing the Average Statistic

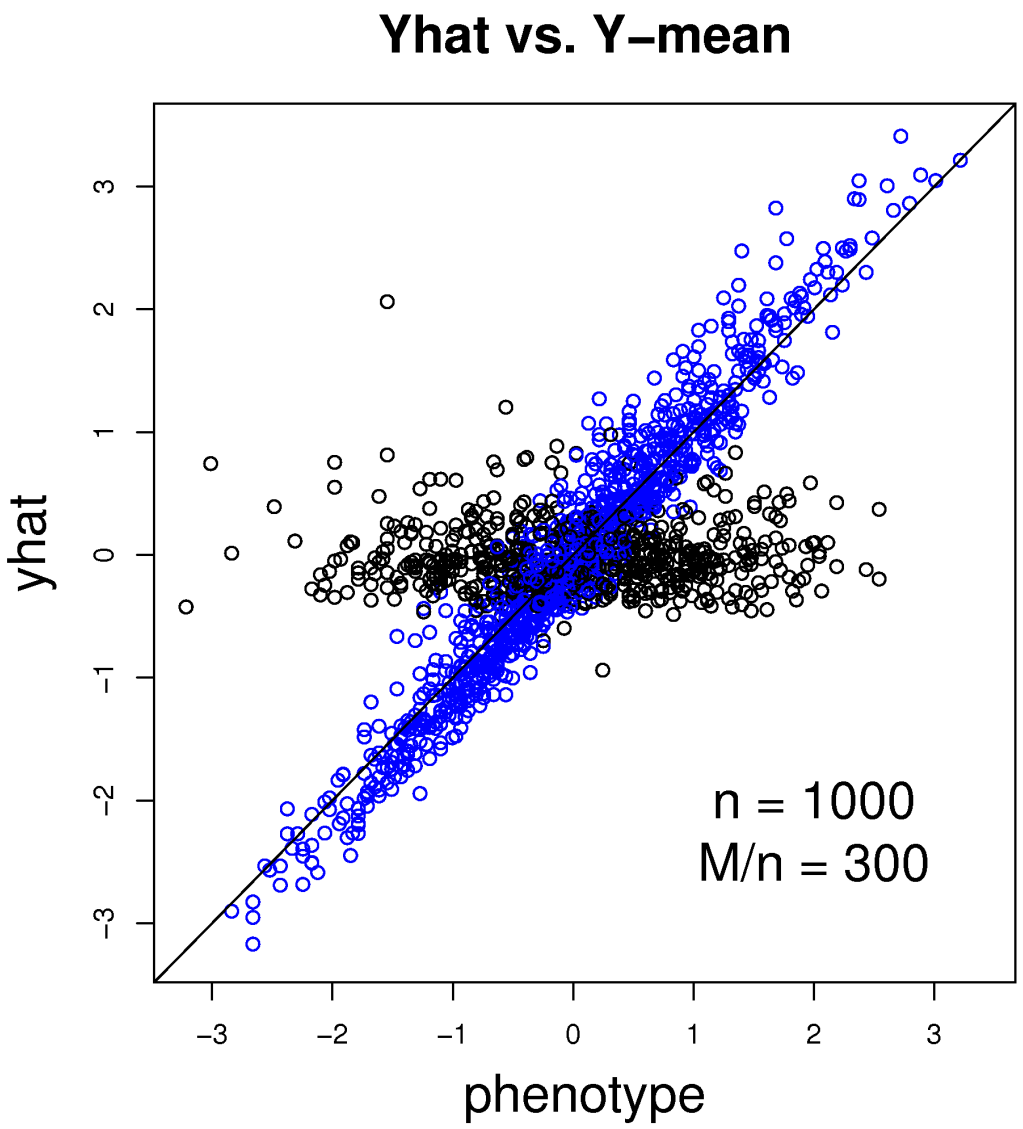
- Dataset from The Genetics of Kidneys in Diabetes Study long-term Type 1 diabetes adults
- phenotype: rank normalized cholesterol level
- Random sample of 1000 individuals
- Remaining 600 used as reference
- using only the 1000 sample ran GWAS

$$\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta_M}$$

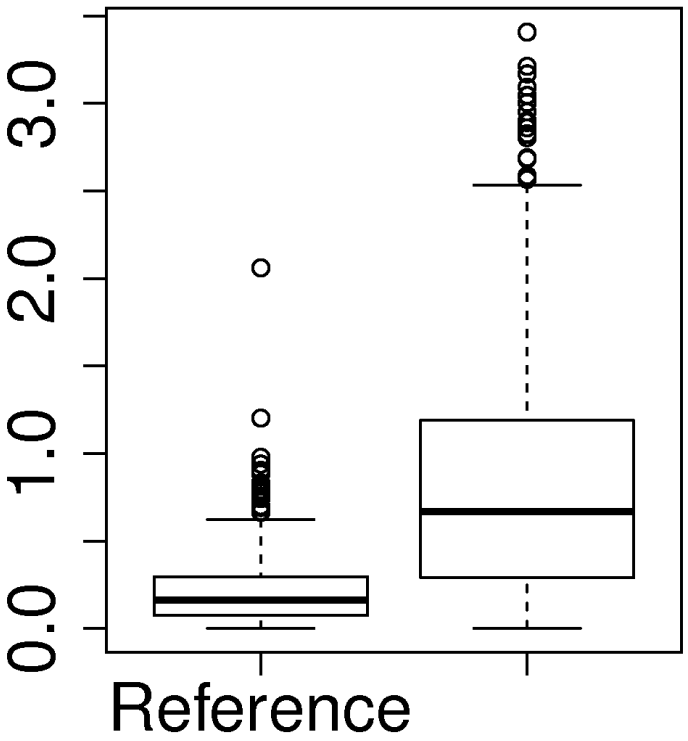
- computed the statistic for all 1600

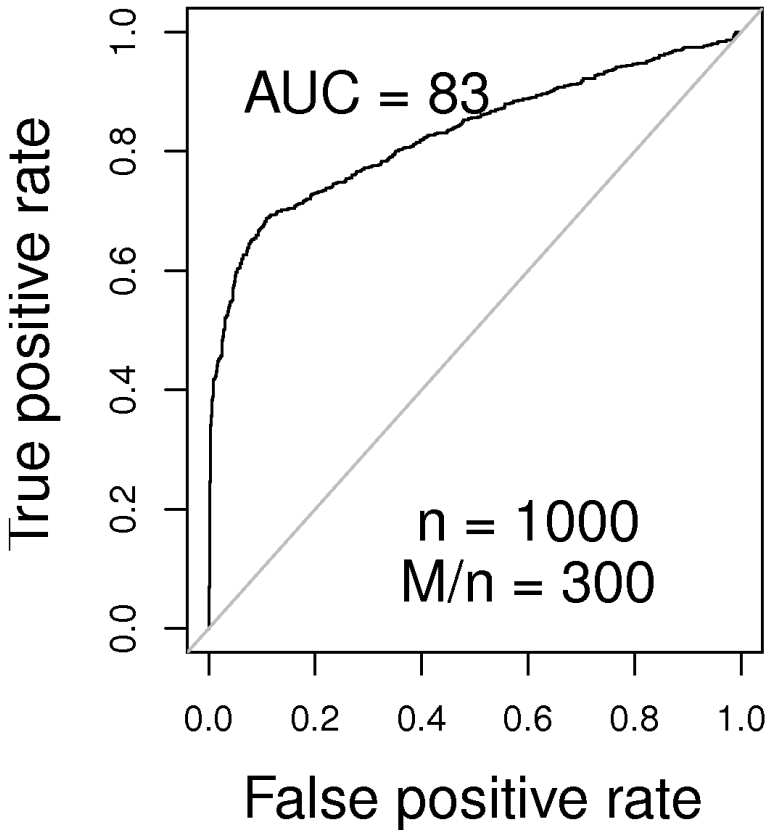
$$\text{Yhat}_I = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}_j X_{I,j}$$

$\hat{Y}$ as predictor of Y - GoKinD data



$\hat{Y}$ distribution and performance - GoKinD data





$\hat{Y}$ Statistics

$$\hat{Y}_I = \frac{n}{M} \sum_{j=1}^{M} \hat{\beta}_j (X_{I,j} - \hat{X}_j)$$

M	# of SNPs
n	# of individuals in the test sample
$X_{I,j}$	allelic dosage of individual I at SNP j
$\hat{X}_j$	estimated mean using the reference group
$\hat{\beta}_j$	estimated β for $Y_i = \alpha_j + \beta_j X_{i,j} + e_i$

Conditional Distribution of $\hat{Y}$

$$
\begin{aligned}
\mathbb{E}\ \hat{Y} \mid X_I, Y_I, \text{in} &\approx (Y_I - \mu) \\
\mathbb{E}\ \hat{Y} \mid X_I, Y_I, \text{out} &\approx O_p\left(\frac{n}{M}\right) \\
\text{Var}(\hat{Y}) \mid X_I, Y_I, \text{in} &\approx \sigma^2 \frac{n}{M} \\
\text{Var}(\hat{Y}) \mid X_I, Y_I, \text{out} &\approx \sigma^2 \frac{n}{M}
\end{aligned}
$$

Power of the Method

$$\text{power} \approx \Phi\left(\frac{|Y_I - \mu|}{\sigma}\sqrt{\frac{M}{n}} - z_{\alpha/2}\right)$$

α: type 1 error

For comparison, when frequencies were known

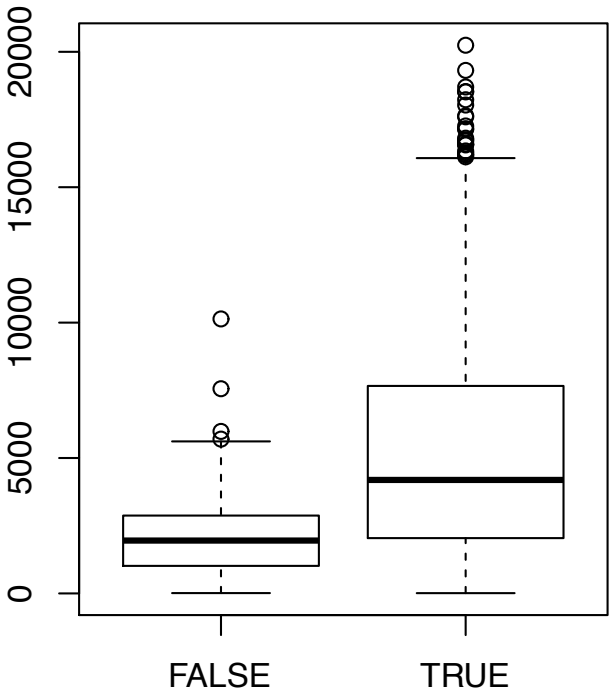
$$\text{power} \approx \Phi\left(\sqrt{\frac{M}{n}} - z_{\alpha}\right)$$

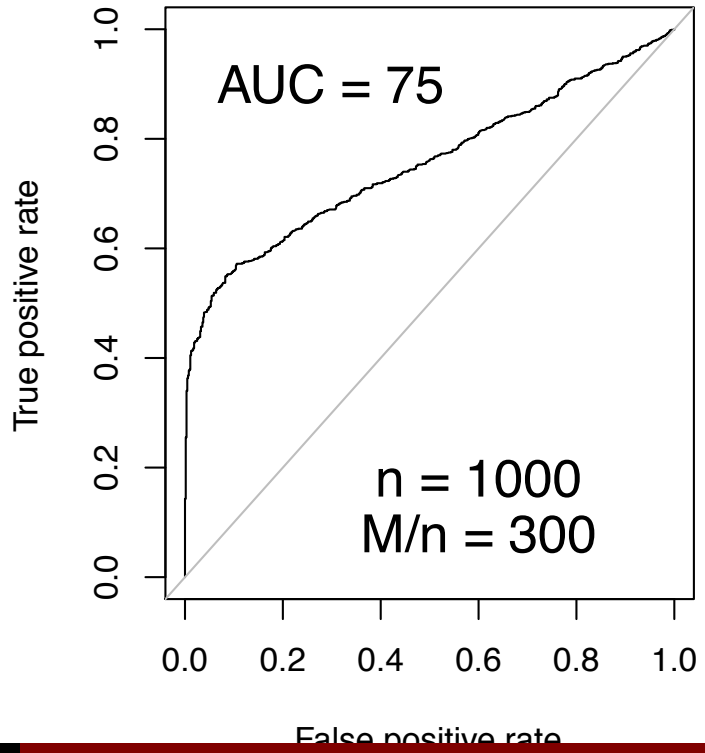
for 5%, 90% power, for $Y_I = \mu + \sigma$

$$13 = \frac{M}{n}$$

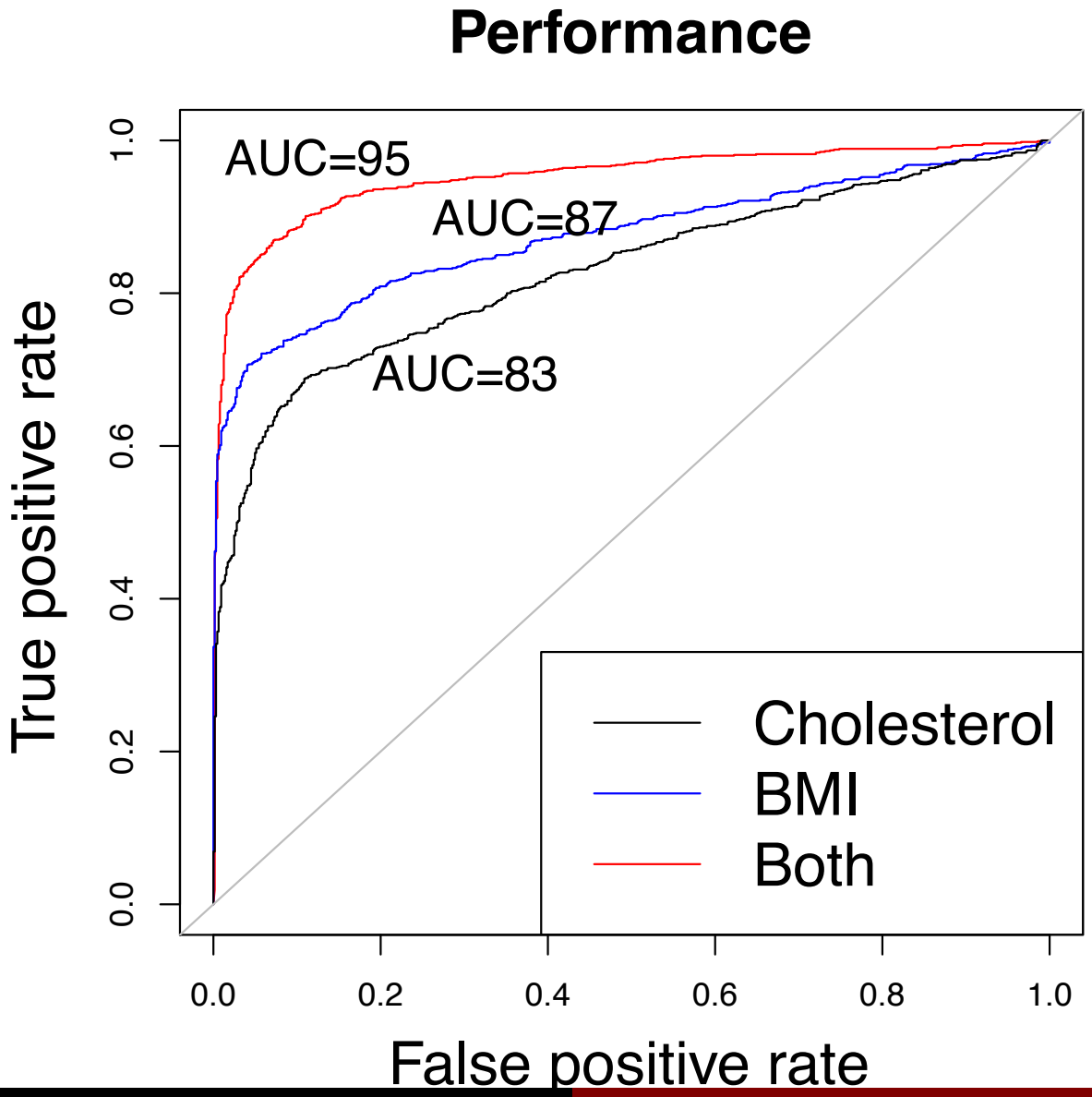
What if Only Direction of Effects Are Known

$$\hat{S} = \sum_{j=1}^{M} \text{sign}(\hat{\beta})\text{sign}(X_{ij} - \hat{X}_j)$$





Performance Improves with Multiple Phenotypes



Summary

- Showed that aggregate results from quantitative GWAS can reveal individual's participation and phenotype
- Computed power of the identification method
- Determined that the direction of effects contains most of the individual's information
- Established that identification becomes more accurate when results from multiple phenotypes are combined
- Thus, there is need to develop data sharing strategies that protect participant's privacy but also facilitate access to data

Acknowledgement

Thank You!

Nancy J. Cox

Eric R. Gamazon

Dan Nicolae

Funding Sources

K12 K12CA139160

GTEx R01 MH101820 and R01 MH090937

PAAR NIH/NIGMS UO1GM61393

Data Sources

GoKinD NIDDK dbGaP Study

Accession: phs000088.v1.p1

IBD NIDDK dbGaP Study

Accession: phs000130.v1.p1

Im, Hae Kyung, Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. 2012. On Sharing Quantitative Trait GWAS Results in an Era of Multiple-Omics Data and the Limits of Genomic Privacy. American Journal of Human Genetics 90 (4): 59198. doi:10.1016/j.ajhg.2012.02.008.